

COMPARATIVE GENOMICS OF MICROBIAL SIGNAL TRANSDUCTION

A Dissertation
Presented to
The Academic Faculty

by

Luke Ulrich

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics

Georgia Institute of Technology
December, 2005

COMPARATIVE GENOMICS OF MICROBIAL SIGNAL TRANSDUCTION

Approved by:

Dr. Igor Zhulin, Advisor
School of Biology
*University of Tennessee – Oak Ridge
National Laboratory*

Dr. Stephen Spiro
School of Biology
Georgia Institute of Technology

Dr. Mark Borodovsky
School of Biology
Georgia Institute of Technology

Dr. Matthew Wolf
College of Computing
Georgia Institute of Technology

Dr. John Kirby
School of Biology
Georgia Institute of Technology

Date Approved: November 10, 2005

I would like to dedicate this dissertation to my father who has constantly encouraged me in my educational pursuits and helped me reach this amazing goal.

ACKNOWLEDGEMENTS

I want to thank Dr. Igor Zhulin for his wisdom, computers, psychotherapy, and Russian jokes. I want to thank my committee members – Dr. Mark Borodovsky, Dr. John Kirby, Dr. Stephen Spiro, and Dr. Matthew Wolf – for their helpful participation. I also want to thank my mother and father for their constant encouragement and support. Thank you Megan for your awesome friendship, the many meals, and help. Thank you Orris for your comic relief and the many lunches we shared as well as your advice on all matters.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF SYMBOLS AND ABBREVIATIONS	xiv
SUMMARY	xvi
<u>CHAPTER</u>	
1 INTRODUCTION	1
Microbial Signal Transduction	1
Impact and Significance	1
Design and Architecture	3
Comparative Genomics	9
Genomic Data Growth	10
Sequence, Structure, Function Relationship	12
Pairwise Sequence Alignment	14
Multiple Sequence Alignment	19
Comparative Genomics of Microbial Signal Transduction	22
Signaling Domains	23
Objectives	29
2 GENERAL MATERIALS AND METHODS	30
Databases	30
Sequence Databases	31
Structural Databases	34

Domain Databases	34
Tools	36
Sequence Similarity Search Tools	36
Multiple Sequence Alignment and Phylogenetics	39
Secondary Structure Prediction	44
Domain Architecture Prediction	47
Visualization	48
3 BIOINFORMATICS PLATFORM AND THE MIST DATABASE	50
Introduction	50
Hardware and Software	50
The MiST Database	55
High-throughput Identification of Signal Transduction Proteins	59
Exploratory Knowledge Environment	66
4 ONE-COMPONENT REGULATORS DOMINATE SIGNAL TRANSDUCTION IN PROKARYOTES	77
Introduction	77
What is a One-component System?	78
Detection of Signal Transduction Proteins in Sequenced Genomes	80
One-component Versus Two-component Systems: a Survey of Bacterial and Archaeal Genomes	82
Genome Size, Lifestyle, and Environment Contribute to the Complexity of Signal Transduction	85
One-component Systems as the Primordial Form of Prokaryotic Signal Transduction	87
Conclusions	88
Acknowledgements	89

5	FOUR-HELIX BUNDLE: A UBIQUITOUS SENSORY MODULE IN PROKARYOTIC SIGNAL TRANSDUCTION	90
	Abstract	90
	Introduction	91
	Methods	92
	Results and Discussion	95
6	RESOLVING THE FUNCTION OF CHEMOTAXIS PAS DOMAINS THROUGH PROTEIN SEQUENCE ANALYSIS	103
	Abstract	103
	Introduction	104
	Methods	106
	Results and Discussion	109
	Acknowledgements	117
7	CONCLUSION	118
	APPENDIX A: Supplementary Information for Chapter 4	119
	APPENDIX B: Supplementary Information for Chapter 5	126
	APPENDIX C: Supplementary Information for Chapter 6	156
	APPENDIX D: Publications	168
	REFERENCES	169

LIST OF TABLES

	Page
Table 1.1: Experimentally characterized signal transduction pathways.	3
Table 1.2: Sensor histidine kinases from the genome of <i>Campylobacter jejuni</i> .	25
Table 3.1: Pfam and SMART domains used for identifying signal transduction proteins.	64
Table 6.1: E-values and scores of the last true positive and first true negative from searches of the PAS_Aer and PAS_Che profiles against the non-redundant database (4 January 2005). Scores are given in the parenthesis following the E-value.	108
Table 6.2: Identification of PAS domains in methyl-accepting chemotaxis proteins by HMM domain profiles.	109
Table A.1: Domains and domain categories used to identify signal transduction systems.	119
Table A.2: Distribution of two-component and one-component systems in prokaryotic genomes.	120
Table A.3: Genomic distribution of input and output domains in archaea and bacteria.	124

LIST OF FIGURES

	Page
Figure 1.1:	<p>Prototypical, bacterial two-component system. In response to a particular stimulus detected by its input domain, the sensor kinase undergoes an ATP-dependent autophosphorylation at a conserved histidine residue in its transmitter domain. The cognate response regulator then catalyzes the transfer of this phosphoryl group to a conserved aspartate residue on its receiver domain, which effects an adaptive response via its output domain.</p>
	4
Figure 1.2:	<p>The EnvZ-OmpR two-component system of <i>E. coli</i>. EnvZ senses osmolarity changes and transmits this signal to its cognate response regulator, OmpR, via a phosphotransfer reaction. In response, OmpR regulates the expression of the <i>ompF</i> and <i>ompC</i> genes, which encode the outer membrane porins, OmpF and OmpC, respectively.</p>
	5
Figure 1.3:	<p>The BvgS-BvgA phosphorelay of <i>B. paraptussis</i>. BvgS senses an unknown signal via its PBP and PAS input domains, which results in the phosphorylation of BvgA via a His-Asp-His-Asp phosphorelay reaction. BvgA responds by regulating the expression of biofilm related genes. Vertical, blue bars represent transmembrane regions.</p>
	6
Figure 1.4:	<p>The chemotaxis system of <i>E. coli</i>. An MCP detects a chemical stimulus and transmits this information to the CheA histidine kinase via protein-protein interactions and the adaptor protein, CheW. CheA undergoes an autophosphorylation reaction at a conserved histidine residue, and subsequently releases the phosphoryl group to the specialized response regulator, CheY. In response, CheY regulates the flagella direction by directly binding to the motor apparatus. CheR and CheB reset the MCP enabling it to sense future changes in the chemical concentration. The phosphatase, CheZ, resets CheY. Vertical, blue bars represent transmembrane regions.</p>
	8
Figure 1.5:	<p>Exponential growth of NCBI's GenBank over the past decade. In 2005, GenBank exceeded 100 gigabases of nucleotide data.</p>
	11
Figure 1.6:	<p>Number of completely sequenced genomes over the past decade.</p>
	11
Figure 1.7:	<p>Sequence, structure, function relationship and the role of comparative genomics.</p>
	13

Figure 3.1:	Structure of the Microbial Signal Transduction database, MiST. Table names are in bold type and lines indicate the various relationships between tables. Values above the lines signify the cardinality constraints for each relationship. The labels 'pk' and 'fk' refer to primary key and foreign key, respectively. In some cases, a single column is both a primary and foreign key, in which case only 'pk' is listed. The color of each table denotes a particular database section: light blue – primary genomic data; purple – derived data; orange – results; and green – management.	58
Figure 3.2:	High-throughput identification of signal transduction proteins.	59
Figure 3.3:	Web interface to PSI-BLAST. Input spaces are provided for the query sequence, database to search, and other various options. Users may also load previous searches using this page.	67
Figure 3.4:	Output of results from a web-based PSI-BLAST search. Descriptive information is given about the search including the BLAST version, database size, and number of hits. A graphical output displays hits to the query sequence as horizontal lines and the color of each line indicates its score. The search may be continued additional iterations or the results from this search used in another analysis (e.g. view the domain architectures for the selected sequences). Significant BLAST hits and information about each pairwise alignment are displayed below the graphical overview.	68
Figure 3.5:	Web interface to the Consensus tool. Users upload ClustalW alignments or input the alignment directly into the provided text area. Checkboxes specify the consensus levels to produce.	69
Figure 3.6:	Output of results from the web-based Consensus tool. The alignment is redisplayed with the consensus above and below.	69
Figure 3.7:	Web interface to the Alignment Shader tool. Users upload ClustalW alignments or input the alignment directly into the provided text area. The user then describes the groups of sequences within the alignment, what color to shade them, and at what similarity threshold a column must surpass for it to be shaded.	70
Figure 3.8:	Output of results from the web-based Alignment Shader tool.	70
Figure 3.9:	Entry web page to the MiST database. Available genomes are listed according to their taxonomy. Clicking the individual checkboxes beside each organism name selects them for further investigation. Users may also search by GI number or MiST identifiers for a specific protein.	72

Figure 3.10:	Analysis selection web page for the MiST database. Each of the previously selected organisms may be searched by domain, description, GI, or MiST identifiers. Clicking on an individual organism name displays various information for that particular organism.	73
Figure 3.11:	Organism specific page for the MiST database containing descriptive information about the genome, signal transduction profile, querying options, and lists of one- and two-component proteins found in this organism.	74
Figure 3.12:	Protein and gene web page for the MiST database. Basic annotation and sequence data for a protein and its corresponding gene are displayed along with the predicted domain architecture and genome neighborhood.	76
Figure 4.1:	Two-component and one-component signal transduction. (a) A prototypical two-component signal transduction system contains input (colored red) and output (colored yellow) domains in two different proteins that communicate via a His-Asp phosphotransfer. A one-component system is a protein that contains input and output domains but lacks His-Asp phosphotransfer domains (colored gray and white). (b) Examples of two-component and one-component systems that utilize the same type of input and output domains. Experimentally studied proteins are identified by name, while proteins predicted from genome sequences are identified by their GenBank ID number.	80
Figure 4.2:	Distribution of input and output domains in bacterial and archaeal signal transduction systems. The counts of the twenty-five most abundant input and output domains in bacterial and archaeal one-component and two-component systems are shown. Domain nomenclature is from the curated Pfam-A database (see Appendix A for detailed information).	84
Figure 4.3:	Dependence of the number of one-component and two-component signal transduction systems on the genome size. The plot is in a double logarithmic scale. One hundred forty-five genomes were ranked by size and split into 16 size classes. Each point indicates the average number of genes for one-component or two-component signal-transduction systems in the respective class.	86

- Figure 5.1: Overview of the VISSA process. For each sequence in a multiple alignment, its secondary structure is predicted and both the alignment and structural information stored in an XML document. This XML data is then visualized by coloring/shading the amino acids that correspond to the predicted secondary structure elements. 94
- Figure 5.2: Alignment of representative members of the 4HB_MCP domain. (A) An alignment of thirty members of the 4HB_MCP domain from the seed alignment. Conserved residues and their positions are colored with the ClustalX scheme using Jalview (Clamp, et al., 2004): orange – glycine (G), yellow – proline (P), blue – small and hydrophobic amino acids (A, V, L, I, M, F, W), green – hydroxyl and amine amino acids (S, T, N, Q), red – charged amino acids (D, E, R, K), cyan – histidine (H) and tyrosine (Y). (B) An alignment of the same thirty members illustrating the VISSA visualization. Regions containing predicted alpha helices and beta sheets have the background shaded red and blue, respectively. The intensity of each shade is directly proportional to the confidence of a given prediction – a darker intensity representing a higher confidence. The identifier for each sequence consists of a species abbreviation, the GenBank identifier, and the coordinates of the sequence. 96
- Figure 5.3: Visualization of conserved residues in the 4HB_MCP domain. (A) A sequence logo generated for the multiple alignment of 282 domain sequences obtained in the PSI-BLAST search initiated with the ligand-binding domain of the *E. coli* Tar chemoreceptor. The secondary structure of Tar is shown above the logo. (B) Conserved tyrosine residues in helices 2 and 3 interact to maintain packing of the four-helix bundle. 99
- Figure 5.4: A schematic view of the domain architecture of 4HB_MCP containing proteins. Domain nomenclature is according to Pfam (Bateman, et al., 2004). Protein GenBank identifiers and the species abbreviations are shown. 101

Figure 6.1:	Multiple sequence alignment with the VISSA visualization of the three subfamilies of PAS domains found in methyl-accepting chemotaxis proteins. Thirty representative PAS domains from each subfamily are shown (for the full alignment of 274 sequences, see Appendix C, Figure C.2). The multiple alignment was constructed using the PCMA (Pei, et al., 2003) and ClustalW (Thompson, et al., 1994) programs, visualized using the VISSA protocol (Ulrich and Zhulin, 2005). Regions containing predicted alpha helices and beta strands have the background shaded in red and blue, respectively. The shading intensity is directly proportional to the confidence of a given prediction – a darker intensity representing a higher confidence. The PAS core, helical connector, and beta scaffold structural regions are delineated above the alignment. Structural elements, as defined by Gong <i>et al.</i> (1998), are listed below the alignment.	111
Figure 6.2:	Neighbor-joining tree built from a multiple sequence alignment of 274 PAS domains showing three distinct subfamilies.	112
Figure 6.3:	Sequence logos for three PAS subfamilies. Highlighted positions are strongly conserved (BLOSUM consensus at least 85% or information content greater than 3 bits). Residues unique to a given subfamily are designated with a triangle (▲). Red arrows indicate residues critical for binding FAD by the PAS domain in <i>E. coli</i> Aer and black arrows indicate other residues that are critical for Aer function (Repik, et al., 2000).	112
Figure B.1:	Unedited seed 4HB_MCP alignment with VISSA visualization.	126
Figure B.2:	Edited seed 4HB_MCP alignment with VISSA visualization.	135
Figure B.3:	Complete 4HB_MCP alignment with VISSA visualization.	141
Figure C.1:	Unedited alignment of chemotaxis PAS domains constructed using PCMA and ClustalW, and visualized with VISSA.	156
Figure C.2:	Alignment of chemotaxis PAS domains after manual editing and visualized using VISSA.	162

LIST OF SYMBOLS AND ABBREVIATIONS

BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Amino Acid Substitution Matrices
CGI	Common Gateway Interface
COG	Cluster of Orthologous Groups
DBI	Database Interface
DNA	Deoxyribonucleic Acid
DP	Dynamic programming
DTD	Document Type Definition
EST	Expressed sequence tag
FLOP	Floating-point operation
FTP	File Transfer Protocol
GB	Gigabyte
GHz	Gigahertz
GI	GenBank identifier
HMM	Hidden Markov model
HTH	Helix-turn-helix
HTML	Hypertext Markup Language
k-mer	A contiguous subsequence of length k
LWP	Library for the World-wide-web for Perl
MCP	Methyl-accepting chemotaxis protein
MEGA	Molecular Evolutionary Genetics Analysis
MiST	Microbial Signal Transduction
NCBI	National Center for Biotechnology Information

NR	Non-redundant database
OS	Operating system
PAM	Point Accepted Mutation
PCMA	Profile Consistency Multiple Sequence Alignment
PDB	Protein Data Bank
Pfam	Protein families
PSI-BLAST	Position Specific Iterative Basic Local Alignment Search Tool
PSSM	Position Specific Scoring Matrix
RAID	Redundant Array of Inexpensive Disks
RAM	Random access memory
RDBMS	Relational Database Management System
RefSeq	Reference Sequence database
SGE	Sun Grid Engine
SMART	Simple Modular Architecture Research Tool
SQL	Standard Query Language
T-Coffee	Tree-based Consistency Objective Function for Alignment Evaluation
VMD	Visual Molecular Dynamics
XML	eXtensible Markup Language

SUMMARY

The goal of this research is to translate genomic sequence data into high-quality biological knowledge on microbial signal transduction. Signal transduction pathways control important cellular functions in all organisms including biosynthesis, catabolism, chemotaxis, development, virulence, host-recognition, and antibiotic resistance. Genome sequencing projects have generated a vast amount of data that has not been analyzed in detail either computationally or experimentally. Up to 40% of genes in most microbial genomes lack assigned biological functions. In the foreseeable future, this gap between the number of sequenced genes and the extent of their functional characterization is expected to broaden further. This creates an urgent need for improved computational prediction of biological function. Signal transduction is one of the most problematic areas for current automated genome annotation protocols, because of the high sequence variability of input and output domains and the mosaic domain architecture of signal transduction proteins. Furthermore, microbiologists studying signal transduction do not actively participate in the validation of computational predictions and the annotation process. This research aims at closing the existing gap between genomic and experimental data in the area of signal transduction in the simplest organisms - prokaryotes.

High-throughput genome processing, sophisticated protein sequence analysis, programming, and information management were used to achieve two major advances in the comparative genomics of microbial signal transduction:

Research Advance 1: A bioinformatics platform and the Microbial Signal Transduction (MiST) database were designed and implemented. An integrated and robust bioinformatics platform and the Microbial Signal Transduction database (MiST) were developed, which facilitated the genome-wide analysis of bacterial signal transduction. This platform was successfully used for the high-throughput identification and classification of over 55,000 signal transduction proteins in more than 300 archaeal and bacterial organisms.

Research Advance 2: One-component systems dominate signal transduction in prokaryotes. Two-component systems that link environmental signals to cellular responses are viewed as the primary mode of microbial signal transduction. A comprehensive review of the signal transduction systems encoded in prokaryotic genomes revealed that contrary to this view, the majority of signal transduction systems are one-component systems – a single protein containing both input and output domains but lacking phosphotransfer domains typical of two-component systems. One-component systems are more widely distributed among bacteria and archaea and display a greater diversity of domains than two-component systems.

Additionally, in-depth bioinformatic analyses were performed that further characterized the function of two, input, signaling domains. Transmembrane chemoreceptors in *Escherichia coli* utilize ligand-binding domains for detecting various external signals. Current domain models for these important sensory modules were constructed from unrelated proteins and fail to detect the majority of domain homologs. A more accurate model, the four-helix up and down bundle, was described that represents

a large domain family with representatives from all major classes of prokaryotic signal transduction including histidine kinases, di-guanylate cyclases, and chemotaxis receptors. PAS domains constitute a widespread superfamily of sensory modules that primarily sense light, oxygen, small ligands, and redox potential. Current computational techniques identify PAS domains in thousands of protein sequences; however, they fail to predict the specific function of a given PAS domain. Two hundred seventy-four PAS domains found in association with chemotaxis receptors were classified into distinct subfamilies. Based on experimental evidence, members belonging to the PAS_Aer subfamily are predicted to function as redox sensors. Members of the PAS_Che subfamily are predicted to bind an unknown ligand and represent a good target for experimental studies. This analysis of chemotaxis PAS domains demonstrates that standard bioinformatics approaches may be applied to multifunctional domains in order to further resolve their function.

CHAPTER 1

INTRODUCTION

Microbial Signal Transduction

For an organism to survive and adapt to changing conditions, it must be able to sense and respond appropriately to environmental signals. Any disruption in an organism's ability to interact with its environment usually produces fatal results or an aberrant and stunted lifestyle. For example, a bacterium ceases to function optimally if it is unable to locate a sustainable energy source or detect changes in the surrounding osmolarity. To prevent useless expenditure of energy, pathogens must recognize a suitable host system before releasing virulent toxins. To combat such pathogens, the host must be able to sense these foreign invaders and then respond with the appropriate immune response. The processing of environmental stimuli directly or indirectly influences the majority of cellular activities in all organisms ranging from bacteria to humans. Therefore elucidating the components and mechanisms of how an organism senses and responds to environmental signals enables a much deeper understanding of the organism and how it behaves, interacts, and adapts to its particular environmental niche.

Impact and Significance

The study of microbial signal transduction includes understanding bacterial systems and pathways responsible for sensing and responding to environmental cues. These systems primarily consist of one or more proteins that behave as sensors and/or regulators that detect a particular stimulus and initiate an adaptive cellular response. Such signaling pathways control the majority of cellular functions including transport,

osmoregulation, respiration, metabolism, biosynthesis, chemotaxis, and development (Stock, et al., 2000). Table 1.1 contains a list of several experimentally characterized systems. Furthermore, signal transduction systems control a significant portion of the genome. For example, the ArcB/ArcA signal transduction system in *Escherichia coli* regulates the expression of more than thirty operons in response to changes in redox conditions (Georgellis, et al., 2001). The NtrB/NtrC system controls up to 2% of the *E. coli* genome in response to nitrogen availability (Zimmer, et al., 2000). In addition, signal transduction systems regulate virulence, host recognition, and antibiotic resistance in important human pathogens, such as *Staphylococcus aureus* (Novick and Jiang, 2003), *Streptococcus pneumonia* (Blue and Mitchell, 2003), *Streptococcus pyogenes* (Biswas and Scott, 2003), *Vibrio cholerae* (Krukoni and DiRita, 2003), *Mycobacterium tuberculosis* (Zahrt and Deretic, 2001), and many others. Signal transduction systems have been reported in the genomes of bacteria that are used in biological warfare, namely *Bacillus anthracis* (Read, et al., 2003) and *Yersinia pestis* (Deng, et al., 2002). Proteins involved in these systems of various pathogenic bacteria are attractive targets for antimicrobial drug design (Matsushita and Janda, 2002; Stephenson and Hoch, 2002). Because of its critical role in bacteria, signal transduction heavily impacts basic research and expands into important industries including medicine, agriculture, biological warfare and bioterrorism, and bioremediation.

Table 1.1 Experimentally characterized signal transduction pathways.

	Signaling pathway	Genes	Reference
Transport	Citrate-update	<i>citA, citB</i>	(Reinelt, et al., 2003)
		<i>bctD, bctE</i>	(Antoine, et al., 2005)
	C ₄ -dicarboxylates	<i>dctB, dctD</i>	(Reid and Poole, 1998)
	Iron	<i>ritR</i>	(Ulijasz, et al., 2004)
	Potassium	<i>kdpD, kdpE</i>	(Walderhaug, et al., 1992)
Osmo-regulation	Outer membrane porins	<i>envZ, ompR</i>	(Cai and Inouye, 2002)
		<i>cpxA, cpxR</i>	(Batchelor, et al., 2005)
Respiration	Aerobic	<i>arcB, arcA</i>	(Georgellis, et al., 2001)
	Redox-control	<i>regB, regA</i>	(Elsen, et al., 2004)
Metabolism	Carbon	<i>uhpB, uhpA</i>	(Island and Kadner, 1993)
	Nitrogen	<i>ntrB, ntrC</i>	(Zimmer, et al., 2000)
	Phosphate	<i>phoR, phoB</i>	(Danhorn, et al., 2004)
Chemotaxis	General	<i>cheA, cheY, cheB, cheR, cheZ</i>	(Armitage, 1999)
Development	Sporulation	<i>kinA, kinB, spo0A, spo0B, spo0F, rapA, rapB</i>	(Stephenson and Hoch, 2002)
Virulence	Pneumococcal	<i>pnpR, pnpS</i>	(McCluskey, et al., 2004)
	Biofilm development	<i>bvgA, bvgS</i>	(Mishra, et al., 2005)
Host recognition	Agrobacterium pathogenesis	<i>virA, virG</i>	(Cho and Winans, 2005)
Antibiotic resistance	Vancomycin	<i>vanR, vanS</i>	(Haldimann, et al., 1997)
	Bacitracin	<i>bceR, bceS</i>	(Ohki, et al., 2003)
	Beta-lactams	<i>croR, croS</i>	(Comenge, et al., 2003)

Design and Architecture

Two-component Systems

Common to all bacterial signal transduction systems is the detection of a signal (input) and coupling this with a cellular response (output). The most widely recognized

signaling systems are so-called two-component systems that utilize protein phosphorylation as a fundamental strategy for signaling (Hoch and Silhavy, 1995; Inouye and Dutta, 2003; Stock, et al., 2000). The prototypical two-component system consists of two proteins, a sensor kinase and a response regulator (Figure 1.1). The sensor kinase detects an environmental signal via its input domain(s). This results in an ATP-dependent autophosphorylation of a conserved histidine residue within the transmitter domain. The response regulator then catalyzes the transfer of the phosphoryl group to a conserved aspartate residue within its receiver domain. Phosphorylation of the receiver domain activates the output domain(s), which effects a particular adaptive response, usually regulation of transcription (Parkinson and Kofoed, 1992).

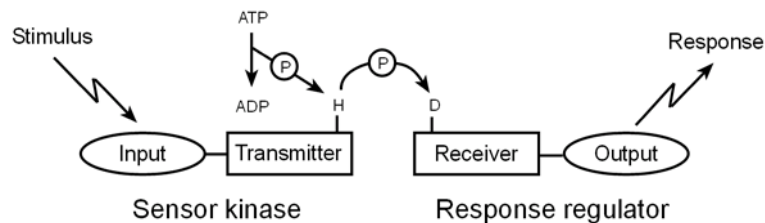


Figure 1.1 Prototypical, bacterial two-component system. In response to a particular stimulus detected by its input domain, the sensor kinase undergoes an ATP-dependent autophosphorylation at a conserved histidine residue in its transmitter domain. The cognate response regulator then catalyzes the transfer of this phosphoryl group to a conserved aspartate residue on its receiver domain, which effects an adaptive response via its output domain.

One of the best-studied systems, the EnvZ-OmpR two-component system of *E. coli* is an example of such a classical signal transduction pathway (Figure 1.2). This pathway regulates the expression of two outer membrane porins, OmpF and OmpC, in response to changes in the osmolarity of the surrounding medium. OmpF and OmpC form channels in the outer membrane that enable the diffusion of small hydrophilic molecules and thus maintain the osmotic pressure of the cell. EnvZ is a membrane bound histidine kinase that senses osmolarity changes via its periplasmic sensory domain. This event triggers the phosphorylation of the conserved histidine in the conserved kinase core of EnvZ. Consequently, this phosphoryl group is transferred to the conserved aspartate residue of the N-terminal, receiver domain of OmpR. This in turn activates the C-terminal, DNA-binding domain of OmpR, which differentially regulates the *ompF* and *ompC* genes and thus appropriately modulates the cell's osmotic pressure (Cai and Inouye, 2002).

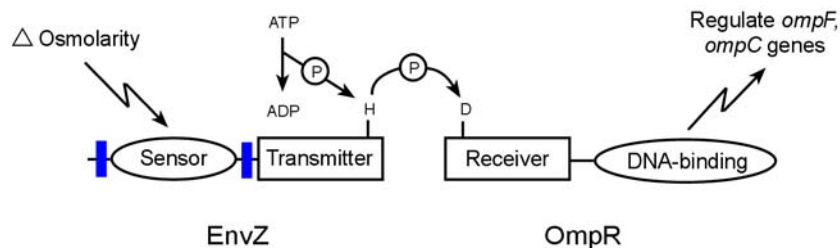


Figure 1.2 The EnvZ-OmpR two-component system of *E. coli*. EnvZ senses osmolarity changes and transmits this signal to its cognate response regulator, OmpR, via a phosphotransfer reaction. In response, OmpR regulates the expression of the *ompF* and *ompC* genes, which encode the outer membrane porins, OmpF and OmpC, respectively. Vertical, blue bars represent transmembrane regions.

Phosphorelays

Phosphorelays extend the basic two-component theme with the inclusion of additional histidine (Hpt) and aspartate containing domains (receiver), which involve multiple phosphotransfer reactions between two or more proteins (Appleby, et al., 1996; Hoch, 2000). The BvgAS two-component system of *Bordetella parapertussis* uses such a phosphorelay in regulating biofilm development (Figure 1.3). In response to an unknown signal (detected by the PBP and PAS input domains), BvgS autophosphorylates at the conserved histidine residue within its transmitter domain. This phosphoryl group then cascades along a His-Asp-His-Asp phosphorelay (i.e. transmitter-receiver-Hpt-receiver), which results in the binding of DNA by the response regulator, BvgA, and regulation of *bvg* related genes (Mishra, et al., 2005). Phosphorelays provide more points of control for fine-tuning the regulation of a system than simpler two-component systems.

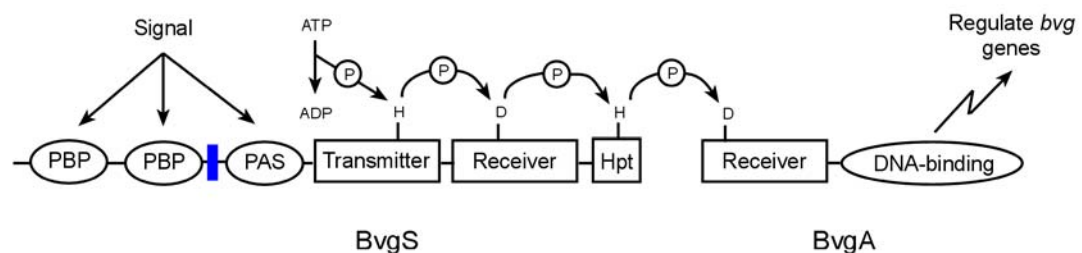


Figure 1.3 The BvgS-BvgA phosphorelay of *B. parapertussis*. BvgS senses an unknown signal via its PBP and PAS input domains, which results in the phosphorylation of BvgA via a His-Asp-His-Asp phosphorelay reaction. BvgA responds by regulating the expression of biofilm related genes. Vertical, blue bars represent transmembrane regions.

Chemotaxis

Chemotaxis is the ability of motile organisms to navigate towards or away from a chemical stimulus and is controlled by one of the most sophisticated signal transduction pathways in microbes (Parkinson, 1993). While there are variations in the chemotaxis pathway(s) between organisms, the *E. coli* chemotaxis pathway is the best-understood and studied system (Figure 1.4). Specialized transmembrane receptors, also known as methyl-accepting chemotaxis proteins (MCPs), respond to changing chemical concentrations and in turn interact with a cytoplasmic histidine kinase, CheA, via the adapter protein, CheW (Boukhvalova, et al., 2002; Falke and Hazelbauer, 2001; Parkinson, 1993). CheA autophosphorylates at its conserved histidine residue and then transfers this phosphoryl group to the receiver domain of its cognate response regulator, CheY. CheY lacks an output domain typical of most response regulators, and instead its receiver domain directly interacts with the flagellar motor to adjust the direction of the flagellar apparatus (Schuster, et al., 2001). CheR and CheB chemically modify the MCP using methyl groups to reset its ability to sense differing chemical concentrations (Djordjevic and Stock, 1998). Finally, CheZ dephosphorylates CheY, restoring its ability to participate with CheA (Zhao, et al., 2002).

Sensor kinases are typically membrane bound with one or more N-terminal sensory domains for detecting a particular stimulus, and a C-terminal transmitter domain containing a conserved kinase core for downstream signaling. Due to the broad range of signals they must sense, input domains comprise a very diverse, highly variable group.

Response regulators are characterized by an N-terminal receiver domain with a conserved aspartate residue used in phosphotransfer reactions and normally contain one or more output domains. These effector domains are more conserved than input domains and tend to have fewer functional roles. The most commonly found output domains are DNA-binding helix-turn-helix (HTH) domains because the predominant adaptive response from response regulators is control of gene expression, which is mediated by interacting with DNA (Parkinson and Kofoed, 1992). Other output responses include RNA anti-termination, regulation of enzymatic activity via adenylate and di-guanylate cyclases, c-di-GMP-phosphodiesterases, phosphohydrolases, and other related domains (Aravind and Koonin, 1998; Chai and Stewart, 1998; Chan, et al., 2004; Jenal, 2004).

In summary, signal transduction represents an incredibly significant cellular activity impacting numerous fields of study and industries. In addition to the functional mechanisms of sensing and responding, other studies document novel signal transduction pathways and the interacting proteins in these pathways (Eguchi and Utsumi, 2005), cross-species variation (von Mering, et al., 2003), structural and biochemical signaling mechanisms (Nioche, et al., 2004; Pappas, et al., 2004), and their environmental implications (Balazsi, et al., 2005).

Comparative Genomics

The nature of experimental characterization of a given gene or protein demands considerable time and resources. Consequently, high-throughput genome sequencing projects uncover novel gene and protein sequences much faster than they may be experimentally analyzed. This underscores the need for computational systems capable of

accurately mapping functional information from known, characterized systems onto the genomic sequence space. This is precisely the goal of comparative genomics, which focuses on the comparison of sequence data from multiple genomes to functionally annotate genes and proteins.

Growth of Genomic Data

Over the past decade, advances in genetics, molecular biology, DNA sequencing techniques and instrumentation have made possible the rapid determination of an organism's DNA – also known as its genome. As a result, numerous high-throughput sequencing initiatives have generated and are continuing to generate, massive quantities of genomic data from all kingdoms of life. This genomic data continues to exponentially increase as evidenced by the growth of GenBank (Benson, et al., 2005), a public repository of nucleotide sequences, which recently exceeded 100 gigabases of nucleotides (Figure 1.5). As of October 2005, genome-sequencing projects have resulted in more than 300 completely sequenced genomes (Figure 1.6) with more than 1600 other genomic projects underway (Bernal, et al., 2001). It is expected that this rate of genomic data production will continue to accelerate with the development of massively parallel sequencing methods that are able to sequence up to twenty-five megabases of DNA in a single four hour period (Rogers and Venter, 2005).

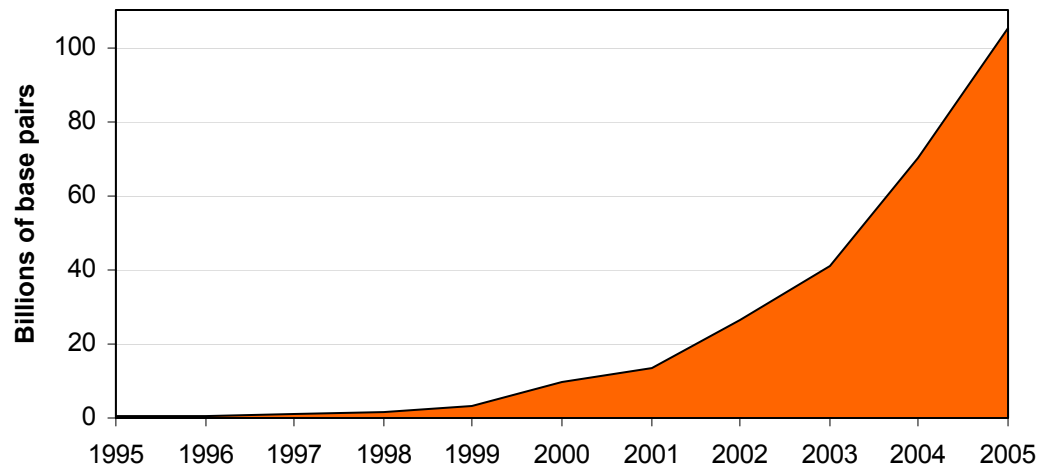


Figure 1.5 Exponential growth of NCBI's GenBank over the past decade. In 2005, GenBank exceeded 100 gigabases of nucleotide data.

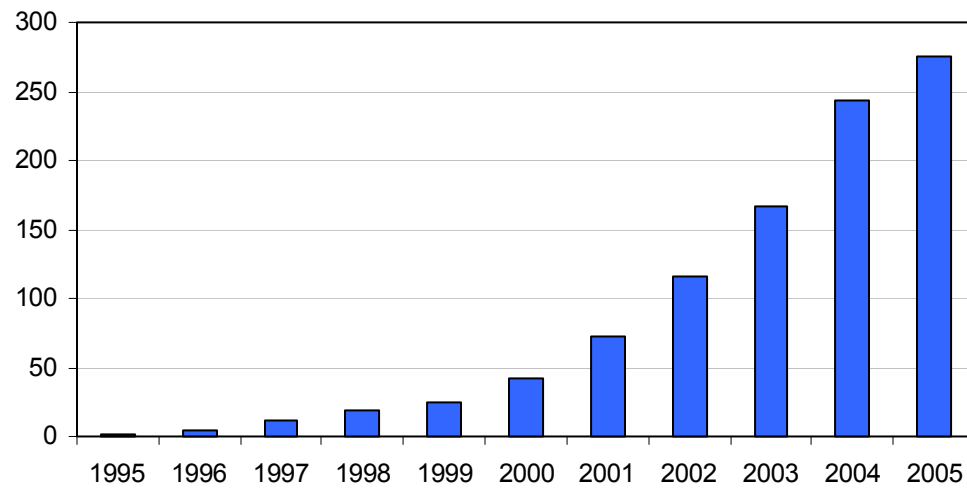


Figure 1.6 Number of completely sequenced genomes over the past decade.

This mountain of genomic data has driven the development of computational tools and methodologies for decoding the stockpiles of DNA into meaningful information. One of the most fundamental and useful tasks entails identifying the genes within a genome. For microbial genomes, computational gene finding tools such as GeneMark (Besemer, et al., 2001) and Glimmer (Delcher, et al., 1999) readily identify 99% of all protein-coding genes. Eukaryotic gene structure is much more complex than prokaryotes due to their exon/intron organization and dissemination across much larger genomes. Therefore, accurate eukaryotic gene delineation remains an active research problem, and sensitive and specific tools are currently being investigated.

Deriving the proteome, the set of all proteins encoded by the genes of a genome, involves determining the protein sequence by translating each gene sequence using the genetic code. This is a trivial matter for microbial genes, but the alternative splicing of eukaryotic genes complicates the accurate prediction of eukaryotic protein sequences. Correctly decoding the gene and protein sequences from microbial DNA marked a major achievement for genomics research, yet was followed by the challenging problem of accurately predicting the function of each putative protein – the goal of comparative genomics.

Sequence, Structure, Function Relationship

Comparative genomics is the science of characterizing the functions of genes and (their derivative) proteins by comparing sequence data from multiple genomes. This process strongly relies on the concept that a protein sequence determines its structure. This protein arrangement presumably determines its function (Figure 1.7). A protein sequence symbolizes the primary structure of a protein as a linear series of amino acids. While this is a convenient representation, the one-dimensional portrayal of a protein

sequence is an artificial construct that poorly models a protein's state *in vivo*. In a biological system, the natural molecular forces act on the linear chain of amino acids, folding the polypeptide chain into a complex 3D structure. The actual shape and spatial configuration of this folded protein determines its specific biological function. The 3D structure of a protein may be determined using techniques such as X-ray crystallography and nuclear magnetic resonance; however, while such methods provide great insight into the actual mechanism and function of a protein, they are quite expensive in terms of equipment, sample preparation, personnel, and finances. Much work has been targeted towards the *ab initio* prediction of a protein's structure from its sequence, but this remains an active research problem and is beyond the scope of this dissertation. Comparative genomics operates under the assumption that two similar (or identical) sequences are likely to produce similar structures with similar biological functions. Presumably, two functionally similar proteins will share equivalent positional and functional residues at the sequence level. Consequently, similar sequences may be assigned a putative function based on comparisons to other similar proteins without knowing the protein's actual structure.

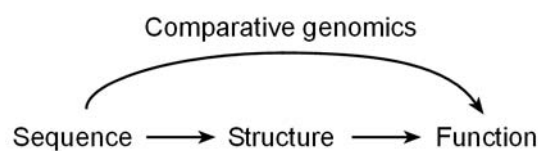


Figure 1.7 Sequence, structure, function relationship and the role of comparative genomics.

Pairwise Sequence Alignment

The alignment of related sequences is fundamental to nearly every comparative genomic study. Sequence alignment is the procedure of searching for common character patterns between sequences in order to reveal their similarity and mutual features. This process involves placing the sequences in rows and shifting the characters (without changing their order) in each sequence such that identical or similar characters are aligned in the same column. Gaps (represented by dashes) may be inserted in order to increase the number of matching characters. An optimal alignment has the maximum possible number of identical and similar characters aligned in each column. A pairwise alignment consists of two sequences aligned to each other, and a multiple alignment comprises three or more simultaneously aligned sequences.

Sequence alignments may be either global or local. A global alignment involves aligning the entire sequences from end to end and is most applicable for highly similar sequences of approximately the same length. A local alignment consists of relatively short subsequences aligned to each other. Local alignments are more appropriate for sequences that share a conserved domain or other functional region that does not extend the entire length of the sequence. In general, most globular proteins consist of multiple functional domains, which occur in a variety of combinations with other domains. Therefore, similarity many times is restricted to relatively short subsequences and protein sequence analysis relies heavily on local alignments at the domain level. This is especially true for analyzing signal transduction proteins, which demonstrate substantial sequence diversity and domain shuffling in their input and output domains.

Scoring

Critical to the alignment process is an appropriate scoring system that numerically defines sequence similarity. Ideally, this will score similar residues higher than dissimilar

residues and a higher score will reflect a more similar biological function. Unfortunately, there is no analytical means for determining a scoring scheme that precisely models all biological systems. Thus, most sequence similarity tools rely on empirically derived schemes for defining similarity scores between every pair of residues. One of the simplest approaches assigns scores based on identical residues – favoring identities and penalizing non-identical residues. While this might suffice for aligning closely related protein sequences, it is well known that sequences with low-identity may have significant similarity and function the same (Koonin and Galperin, 2003). A far superior solution is the use of substitution matrices. A substitution matrix is a 2D probability matrix containing values for every possible pair of residues and numerically represents the odds for which one residue is likely to be substituted by itself or another residue. For mathematical convenience, these odds scores are typically converted into logarithms and thus called log-odds scores. Several different substitution matrices have been derived including ones that attempt to model the chemical, functional, charge, and structural properties of the various amino acids (Karlin and Ghandour, 1985) or based solely on structural similarities (Feng, et al., 1985); however, the two most popular matrices are the PAM (Dayhoff, et al., 1978) and BLOSUM (Henikoff and Henikoff, 1992) matrices, which are based on observed counts of amino acid changes within closely related sequences.

Substitution Matrices

One of the first substitution matrices to take on widespread use was Dayhoff's PAM (Point Accepted Mutation) matrices (Dayhoff, et al., 1978). These matrices are based on 1572 substitutions found within 71 groups of protein sequences that were at least 85% similar with the intent to capture the relative mutability of each amino acid over a given period of evolutionary time. The underlying assumption when building these matrices is that the rate of mutability may be extrapolated to produce additional matrices

that more accurately weight the likelihood of an amino acid change over longer evolutionary times. The PAM-1 matrix is equivalent to 1% divergence of a protein or one amino acid difference per 100 residues. Because the mutability of each amino acid is assumed to be independent and modeled with a Markov chain, the PAM-1 matrix may be multiplied by itself N times to produce a PAM-N matrix. For example, the PAM-250 would be the PAM-1 matrix extrapolated 250 times and would score pairs of residues based on 250 substitutions per 100 residues. Despite this large number of substitutions, sequences at this level of divergence are still about 20% similar.

Unlike the PAM matrices, the BLOSUM (Blocks Amino Acid Substitution Matrices) matrices are not based on an explicit evolutionary model, but are derived from substitution counts observed in approximately 2000 conserved amino acid patterns known as blocks. These blocks were taken from the BLOCKS database (Henikoff and Henikoff, 1991), which at the time represented more than 500 distinct protein families classified by their biochemical function. Each block is an ungapped alignment of an amino acid pattern shared by members of a particular protein family. The number of amino acid pairs was counted from each block and converted into log-odds substitution scores. Several distinct matrices were built from blocks with differing levels of identity. For example, one of the most popular matrices, BLOSUM62, was constructed from sequences that were at least 62% identical (Henikoff and Henikoff, 1992).

Methods

The primary method for constructing pairwise alignments ordinarily includes some form of dynamic programming (DP). DP was originally used for making global alignments (Needleman and Wunsch, 1970) and later modified to produce local alignments (Smith and Waterman, 1981). DP compares every pair of residues between two sequences and builds a matrix of substitution scores that represents all possible alignments. Matches and mismatches are scored using a substitution matrix and gaps are

penalized. The algorithm then produces the optimal alignment by tracing through the matrix and identifying and joining the highest scoring segments. DP has been proven to produce the mathematically optimal alignment (Lipman, et al., 1989), yet its execution requires computation and memory proportional to the square or cube of the sequence lengths. Despite the appeal of finding the optimal alignment, the slower performance of DP limits its application.

Significance

The development of statistical methods for estimating the significance of a pairwise alignment marked a major advance in protein sequence analysis. Determining and confirming the biological importance of a particular alignment is a difficult task, especially for proteins with low sequence similarity. Searches against sequence databases containing millions of sequences exacerbate this problem by producing numerous alignments. Thus, modern tools first estimate the significance of each alignment and filter out hits that fall below a user-specified threshold. The significance of a local alignment score may be determined given the distribution of alignment scores between completely random and unrelated sequences that have similar length and composition to the input sequences. Such random alignment scores follow a Gumbel extreme value distribution. If a particular alignment score is much greater than that expected by chance as determined by comparing to this extreme value distribution, the alignment is considered to be significant. The probability of an alignment occurring by chance, also known as the E-value, is determined using the Karlin-Altschul equation (Altschul, et al., 1994; Altschul and Gish, 1996; Altschul, et al., 1990). Low E-values indicate a significant alignment and *vice versa*. The statistical mathematics and complexities underlying the derivation of the Karlin-Altschul equation and other related equations are outside the scope of this dissertation.

Database searches

Popularized by its rapid searching method and ability to statistically rank pairwise alignments, the Basic Local Alignment Search Tool (BLAST) quickly became the *de facto* standard for performing sequence similarity searches. Unlike DP, BLAST does not find the optimal alignment, but rather approximates the optimal alignment using a heuristic search method. Specifically, BLAST locates all common *k*-mers, a contiguous subsequence of length *k*, between two sequences that score above a certain threshold. All the highest scoring segments are then extended and linked using DP and the resulting alignments estimated for their significance using Karlin-Altschul statistics. BLAST displays only those pairwise alignments scoring above a user-specified E-value. By restricting DP to sequence segments with promising similarity, BLAST bypasses searching the entire sequence space and thus overcomes the speed and memory limitations of DP. BLAST made possible the rapid querying of large sequence databases for homologous sequences (Altschul, et al., 1990). Consequently, the first clue to determining a novel protein's function often involves "BLASTing" its sequence against the SwissProt (Boeckmann, et al., 2003), non-redundant, and/or other protein databases. Furthermore, most automated annotation pipelines attempt to assign a particular function to each protein of a new proteome using BLAST (Fraser, et al., 2000).

Position Specific Iterative (PSI)-BLAST is a sensitive sequence similarity search tool that uses an iterative searching method and unique scoring scheme to detect weakly related homologs (Altschul, et al., 1997). The first iteration is equivalent to a normal BLAST search, and PSI-BLAST scores alignments using a standard substitution matrix (e.g. BLOSUM62). After each additional iteration, PSI-BLAST scores all pairwise alignments using a position-specific scoring matrix (PSSM) that was derived from a multiple alignment of the significant hits from the previous iteration (see below). The search process continues until convergence or when no new homologs are retrieved. PSI-BLAST is quite sensitive because the scoring matrix dynamically adapts to the observed amino acid substitution rate specific to each search and information from multiple

sequences is used rather than a single sequence. However, PSI-BLAST often requires human input in order to prevent unrelated hits from “corrupting” the scoring matrix and incorrectly biasing the search toward false positives. The ability to recognize weak sequence patterns makes PSI-BLAST a critical tool for comprehensively identifying all the members of a protein family or instances of a functional domain. Because the input and output domains of signal transduction proteins display low sequence similarity, their complete identification mandates the use of PSI-BLAST.

Multiple Sequence Alignment

Multiple sequence alignments provide an invaluable source of information into characterizing a group of related sequences. Alignments of protein sequences may be used to identify conserved amino acid patterns (Bork, et al., 1994), define functional and structural domains, identify novel members of a protein family (Eddy, 1998), build phylogenetic trees (Saitou and Nei, 1987), structure prediction (Rost, et al., 1994), and assist in experimental investigation. Despite their utility, constructing high-quality multiple alignments for divergent sequences (of less than 30% identity) is a laborious process often requiring the combination of computational tools and manual editing. Oftentimes in characterizing a functional domain, hundreds and sometimes thousands of sequences need to be aligned. In theory, producing an optimal multiple alignment is possible using a generalized version of the DP pairwise alignment algorithm; however, in practice, the space and time complexity of DP limits this approach to a small number of sequences (Lipman, et al., 1989). As a result, multiple alignment algorithms heuristically approximate the optimal alignment using either a progressive, consistency-based, iterative, or structure-based alignment strategy. Most tools adopt a hybrid approach that combines one or more of these alignment strategies.

Methods

Progressive alignment performs a pairwise alignment comparison of all sequences to each other using DP and then progressively aligns the most similar sequences or groups of sequences until no more sequences remain. The major disadvantage of this method is that the outcome strongly depends upon the order in which the sequences are aligned and propagates errors to any remaining sequences that have yet to be aligned. Even so, progressive alignment tends to execute rapidly, handle large numbers of sequences, and provide a relatively acceptable alignment that may serve as the starting point for further manual modification. The most popular tool for progressively aligning either DNA or protein sequences is the ClustalW program (Thompson, et al., 1994). Consistency-based measures use information from the pairwise alignments of all input sequences to determine the order of a progressive alignment and to appropriately weight the pairing of any two residues. T-Coffee (Notredame, et al., 2000) demonstrates this procedure and works well with less conserved sequences at the cost of a significantly longer execution time and memory footprint. Iterative strategies first create a draft alignment and then continue to refine this alignment until it surpasses a certain threshold. MUSCLE iteratively aligns large numbers of sequences in a reasonable time and the authors claim comparable results to T-Coffee (Edgar, 2004).

One of the goals of creating a multiple alignment for a particular protein domain is to align the residues with structural equivalence (i.e. each column will contain residues that support a particular structural feature such as an alpha helix or beta strand). Because a multiple alignment is expected to model the common structure of a group of related proteins, one recent method (Zhou and Zhou, 2005) for aligning proteins first predicts the secondary structure of each protein using PSIPRED (Jones, 1999) and then incorporates this structural information into the alignment process. This method demonstrated a 7-15% improvement in alignment accuracy for divergent (less than 30% identical) sequences.

Protein Domains

Multiple sequence alignments serve as the basis for representing and modeling protein domains. There are two types of protein domains: structural and homologous. A structural domain is the unit of a protein that can independently fold into a distinct, functional, 3D structure. Such structural units have identifiable patterns of amino acid conservation in sequence-space that are recognizable by computational protein sequence analysis.

Homologous protein domains are based on sequence characteristics from a group of closely related sequences that have been multiply aligned. Presumably, the multiple alignment reveals which amino acids are permitted in a given column and reflects the amino acid's structural function.

Sound statistical principles confirm that protein domains can be robustly represented by profile hidden Markov models (HMMs) (Eddy, 1998). A typical HMM is a linear series of match, insert, and delete states that are analogous to the similar characters and gaps of multiple sequence alignments. Associated with each state is a set of transition probabilities and a set of character probabilities. The former represents the likelihood of changing to a different state (e.g. one amino acid substituting or deleting for another), and the latter represents the probability of each character (in this case, amino acids) occurring in this state. These probabilities are calculated from a training set of homologous sequences that have been aligned using programs such as ClustalW (Thompson, et al., 1994) or MUSCLE (Edgar, 2004). After the training process, the HMM contains information regarding the conserved primary structure of the homologous domain sequences. In addition, each HMM may have user-supplied thresholds incorporated into its data structure which specify statistical cutoffs for delineating true members of a particular domain. Scanning a sequence database with each HMM may then identify novel domain homologs. Collections of such HMMs are stored in the primary domain databases, Pfam (Bateman, et al., 2004) and SMART (Letunic, et al., 2004).

Apart from identical sequences or experimental evidence, it is not ultimately possible to predict with certainty that two similar proteins will function the same. Therefore, while sequence similarity and comparisons form the core of comparative genomics, it is the integration and synergy of several independent lines of genomic (and experimental, when available) evidence that substantiates a functional prediction. Additional methods that were helpful in this research include the following: sequence similarity trees, sequence logos, and genome neighborhood analysis. Sequence similarity trees are useful for depicting the relationship between several sequences using a binary tree diagram. Because similar sequences cluster together, these trees visualize the distribution of related sequences and may reveal potential subfamilies with distinct functions (Saitou and Nei, 1987). Sequence logos graphically visualize the conservation of each position in a multiple alignment. This highlights conserved residues that may play an important functional role (Crooks, et al., 2004). Finally, adjacent genes are often co-expressed suggesting a common cellular role and signal transduction genes oftentimes are located beside the operons that they regulate. Therefore, the genomic neighbors for a particular gene or protein of interest may reveal significant clues as to a protein's function (Dandekar, et al., 1998; Overbeek, et al., 1999).

Comparative Genomics of Microbial Signal Transduction

The explosion of genomic data and steady increase of uncharacterized proteins necessitates the development of improved computational techniques for accurate prediction of protein function. Current automated annotation procedures often fail to adequately describe signal transduction proteins due to the high sequence variability of input and output domains and their mosaic domain architecture. Most signaling proteins in public databases are poorly annotated as “putative two-component sensor” or “two-component response regulator” partly because they are based on crude sequence similarity searches that reveal similarity to the conserved transmitter and receiver

domains. Furthermore, there is no comprehensive electronic resource or database for microbial signal transduction. The study of signal transduction could benefit greatly from advanced comparative genomic studies due to the numerous experimentally characterized systems and the lack of sufficient methods for automatically producing high quality annotations of signal transduction proteins.







At this time, prokaryotes are the best model for using computational methods for the analysis of signal transduction systems. First, signal transduction pathways are much simpler in prokaryotes (typically, one to four protein components) than in eukaryotes (branched multi-protein cascades). Second, prokaryotic genes are often grouped together in operons, and signal transduction proteins are often associated with operons that they regulate. This makes genome context analysis a useful method for revealing the function of prokaryotic proteins. Third, more than 300 prokaryotic genomes are currently available (versus twenty-eight eukaryotic genomes) that span much longer evolutionary distances (which is critical for comparative analysis) than eukaryotes,. Fourth, prokaryotic gene-finding tools are considerably more sensitive and specific than eukaryotic gene-finding tools. This provides a much more comprehensive and accurate genetic record when working with multiple genomes. The smaller size of microbial genomes facilitates the management of genomic data. Finally, the well-established signal transduction community provides a source of experimental input – a critical link in ascertaining the validity and application of computational predictions.

Signaling Domains

The combination and type of signaling domains in a signal transduction protein determines its specific function. Therefore, reliable prediction of function for a signal transduction protein can only be derived from its complete domain architecture. The comparative genomics of microbial signal transduction largely comprises characterization of these signaling domains and interpretation of the cellular role of signal transduction

pathways at the domain level. Signaling domains belong to one of four classes of domains: input, transmitter, receiver, and output. They are principally modeled using HMMs (see above), which enables their rapid identification in protein sequences using the HMMER program. Preliminary analysis demonstrated that the majority of signal transduction proteins have partially characterized domain architectures with one or more regions where no domains are identified. The regions in prokaryotic protein sequences that might contain a domain are typically 80-100 amino acids long (Koonin and Galperin, 2003). Therefore, regions of eighty amino acids or longer that lack any statistically significant automated domain prediction, putatively contain an unknown or undetected domain. Examples of such undetected domains are shown in Table 1.2, which contains data for all six sensor histidine kinases encoded in the genome of the human pathogen *Campylobacter jejuni*. Each of these sensor proteins contains at least one undetected domain in their N-terminal input region, as revealed by PSI-BLAST searches against the non-redundant database. Nonetheless, these domains are not detected in automated searches against SMART (Letunic, et al., 2004) and Pfam (Bateman, et al., 2004). Therefore sensory function for any of these proteins was not predicted, and the proteins are currently incompletely and inconsistently annotated.

Table 1.2 Sensor histidine kinases from the genome of *Campylobacter jejuni*.

GI number	Domain architecture	Annotation (NCBI)
15792131		signal transduction histidine kinase
15792219		putative sensory transduction histidine kinase
15792546		putative two-component sensor
15792550		putative two-component sensor
15792586		two-component sensor (histidine kinase)
15792807		putative two-component sensor

Transmitter and Receivers

The homologous domains that comprise transmitter and receiver modules of signal transduction proteins are well conserved in sequence. Both SMART and Pfam contain high-quality HMMs for these domains. Thus, histidine kinases and response regulators are easily detectable by HMMER and related programs (Eddy, 1998). Scanning against Pfam and SMART has become a standard procedure for whole-genome annotation, and this has resulted in the successful identification of two-component signal transduction systems in completely sequenced microbial genomes (Grebe and Stock, 1999). Detecting transmitter and receiver modules in signaling proteins is an important and necessary step, but it does not provide answers to the most important biological questions: (1) What signal is detected by a given signal transduction system? and (2) What adaptive response does it produce? The detection of the input and output modules is necessary to provide the answers to these fundamental questions.

Input and Output Domains

Unlike transmitter and receiver domains, input and output modules are significantly less conserved in sequence. Because input domains must respond to such a broad range of signals, they are particularly variable and more difficult to identify by computational means. Relatively few input domains specific to signal transduction have been characterized, but they include the following: PAS (Ponting and Aravind, 1997; Zhulin and Taylor, 1997), GAF (Aravind and Ponting, 1997), Cache (Anantharaman and Aravind, 2000), CHASE (Anantharaman and Aravind, 2001; Mougél and Zhulin, 2001), CHASE2 through CHASE6 (Zhulin, et al., 2003) and NIT (Shu, et al., 2003).

Output domains are more highly conserved than input domains. The most commonly found output domains are DNA-binding helix-turn-helix (HTH) domains because the predominant adaptive response from response regulators is control of gene expression, which is mediated by binding to DNA. Several novel output domains have been recently described in response regulators, which implicate these systems in other types of control, such as the regulation of enzyme activity. These include adenylate and diguanylate cyclases, c-di-GMP-phosphodiesterase, phosphohydrolase, and other related domains (Galperin, et al., 2001; Nikolskaya and Galperin, 2002; Pei and Grishin, 2001; Shu and Zhulin, 2002).

Computational Characterization of Signaling Domains

All of the previously mentioned input and output domains were detected by using the Position-Specific-Iterative (PSI) BLAST program (Altschul, et al., 1997), a sensitive sequence-similarity search tool that requires manual control of parameters such as the inclusion threshold, low-complexity filter, composition-based statistics, and exhaustive searches seeded with any new homologs detected after each iteration (Altschul and Koonin, 1998; Koonin and Galperin, 2003; Schaffer, et al., 2001). These factors make PSI-BLAST a time-consuming and lengthy procedure. Furthermore, the follow-up work

describing a particular domain family involves construction of multiple alignments, identification of conserved residues and their contribution to a known or predicted structure, and determination of the domain architectures of related proteins. Thus, identification and characterization of novel domains in signal transduction proteins is a laborious procedure involving a variety of bioinformatics tools applied on a case-by-case basis.

Examples of Signaling Domains

PAS

The PAS domain is one of the most ubiquitous and important input domains in signal transduction (Taylor and Zhulin, 1999). PAS domains comprise a widespread domain superfamily, which are responsible for detecting light, oxygen, and redox potential. The computational characterization of PAS (Zhulin and Taylor, 1997) involved exhaustive PSI-BLAST searches using sequence data from the oxygen receptor, Aer, of *E. coli* (Rebbapragada, et al., 1997). Both Pfam and SMART contain models for the PAS domain and local HMM searches with this domain profile reveal more than 3500 signal transduction proteins containing this domain (October, 2005).

CHASE and NIT

Two other input domains recently described are the CHASE and NIT domains. Found within both prokaryotes and eukaryotes, the extracellular CHASE domain is predicted to bind small ligands (Mougel and Zhulin, 2001). This novel domain was characterized using exhaustive PSI-BLAST searches combined with multiple alignment construction, secondary structure prediction, identification of transmembrane regions in protein sequences, and analysis of amino acid conservation patterns. Essentially the same computational tools and methods were used for the identification of the NIT domain (Shu, et al., 2003). NIT containing proteins are predicted to carry out a very specific

function: detection of nitrate and nitrite in the extracellular environment. This prediction was made by extrapolating detailed biochemical data available for one of the homologs (Chai and Stewart, 1998).

ANTAR

Based on the analysis of an available three-dimensional structure, PSI-BLAST searches and other computational tools, Shu and Zhulin (2002) described ANTAR, an RNA-binding, output domain that is present exclusively in signal transduction proteins. The experimentally derived structure and function of this domain permitted the characterization of this domain despite only 5% sequence identity among sequence members. ANTAR carries out an important step in transcription anti-termination; therefore its identification in several human pathogens, such as *Mycobacterium tuberculosis*, *Listeria monocytogenes*, *Burkholderia fungorum*, *Klebsiella pneumoniae*, provides important information that may be used in antimicrobial drug design.

Although in many instances the exact sensory or regulatory capabilities of a novel domain cannot be predicted directly, solving the complete domain architecture of proteins that contain this domain provides important clues about the biological function of the protein. Recently five novel input domains (CHASE2 through CHASE6) of unknown ligand-binding capabilities were identified (Zhulin, et al., 2003). By defining the complete domain architecture of related proteins containing these domains it was demonstrated that in bacteria similar signals can be transmitted via different signal transduction pathways to different regulatory circuits. The pervasiveness of a particular input domain in receptors from diverse regulatory networks in a given organism suggests the importance of a signal recognized by these receptors. These results provide evidence that refined biological function for signal transduction proteins can be predicted by solving the complete domain architectures.

Objectives

This research aims at reducing or closing the existing gap between genomic and experimental data in the area of signal transduction in the simplest organisms - prokaryotes. To achieve this goal, we pursued three major objectives. First, we built a customized and integrated bioinformatics platform to facilitate the genome-wide analysis of signaling systems. Second, using this platform, we sought to identify and analyze signal transduction proteins to investigate overarching trends throughout microbial signal transduction. Thirdly, we performed in-depth characterization of two sensory domains to provide a deeper understanding of specific signaling systems and to extrapolate this information onto existing and novel genomes.

CHAPTER 2

GENERAL MATERIALS AND METHODS

Since the inception of bioinformatics research, numerous tools and databases have been designed and they vary quite dramatically in complexity, scope, and application. Because bioinformatics research largely relies on comparisons, countless databases (Galperin, 2005) have been developed to manage widely diverse data sets. These range from raw collections of sequence data to highly specialized databases such as the EcoCyc database (Keseler, et al., 2005), which is specifically designed for managing information about the gram-negative bacterium, *E. coli*. In order to gain the maximal utility from these databases, many software programs have been designed to derive novel, additional information from primary data (such as a gene locus or protein sequence). The resulting output which itself might be incorporated into a secondary derivative database. The integrated use of multiple approaches requires insight and familiarity with the databases and their associated tools, so that they be efficiently applied to a particular biological problem. In this chapter, relevant bioinformatic tools and databases are described that aided in the research presented in this dissertation.

Databases

Three major types of databases discussed in this section:

- 1) Sequence databases that are simply repositories of either DNA or protein sequences with minimal annotation,
- 2) Structural databases containing biological macromolecular information, and
- 3) Domain databases that characterize various protein families.

Sequence Databases

GenBank

GenBank is a public database of nucleotide (DNA) sequences maintained by the National Center for Biotechnology Information (NCBI) as part of the National Institutes of Health. GenBank currently contains over 100 gigabases (1×10^{11}) of nucleotide data sequenced from more than 165,000 organisms

(http://www.nlm.nih.gov/news/press_releases/dna_rna_100_gig.html, August 2005).

GenBank continues to grow exponentially due to improved methods for sequencing new organisms and sequences from nearly 2000 new species are added every month.

Traditionally, GenBank was partitioned based on taxonomy. Newer divisions such as the whole genome shotgun division, have been created to accommodate data from high-throughput sequencing initiatives. GenBank participates in the International Nucleotide Sequence Database Collaboration, which consists of GenBank, the European Molecular Biology Laboratory Data Library, and the DNA Data Bank of Japan. As part of this worldwide effort, these organizations exchange nucleotide data on a daily basis.

Each nucleotide sequence receives a stable identifier called an accession number and another unique identifier known as the GenBank Identifier (GI). While the accession number does not change, any updates to the sequence results in a new GI number and increments the record version. Incorporated into the annotation of each record are protein translations, features, and references. There is significant redundancy within GenBank owing to its comprehensive collection of DNA from multiple sources. GenBank is publicly available and may be accessed via the internet with services such as Entrez found at NCBI's site (<http://www.ncbi.nlm.nih.gov/Entrez/>) or downloaded from <ftp://ftp.ncbi.nih.gov> (Benson, et al., 2005).

NR – Non-Redundant Database

The NCBI non-redundant database (NR) is a comprehensive, non-redundant set of all protein sequences contained in GenPept, SwissProt, the Protein Information Resource, the Protein Data Bank, and the Reference Sequence database. In practice, NR contains some redundancy due to sequencing errors and minor differences between closely related proteins; however, it represents a good approximation at compiling identical sequences into a single entry. Currently, NR contains approximately three million proteins and represents the most complete “snapshot” of all available protein sequences. NR may be downloaded from NCBI at <ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz>.

RefSeq – Reference Sequence Database

The NCBI Reference Sequence (RefSeq) database is a public, non-redundant database of genomic data, transcripts, and protein sequence information with the goal of providing an updated synthesis of both primary and secondary information for each sequence record. RefSeq is characterized by three features: 1) it is a curated, non-redundant collection of sequences; 2) it contains explicitly linked protein and nucleotide records to external sources of information; and 3) it represents significant taxonomic diversity with data from viruses, prokaryotes, and eukaryotes. RefSeq contains over one million protein sequences from more than 2,400 organisms. The annotation of each sequence is distinct from the original submission to GenBank with information gathered from multiple sources including: the original GenBank record, automated computational analyses, collaborators, manual curation, and user feedback. Similar to GenBank, each sequence is given a stable accession number and unique GI number. The RefSeq database is available for download at <ftp://ftp.ncbi.nih.gov/refseq/release> (Pruitt, et al., 2005).

SwissProt and TrEMBL

The SwissProt database is a high-quality collection of manually annotated, protein sequences and strives to provide state-of-the-art information for each sequence by combining experimental results and computational analysis. At a minimum, each record in SwissProt must contain the amino acid sequence, the protein name, its taxonomic classification, and a literature reference. Additionally, SwissProt contains extensive cross-references to external databases and each record follows a strict naming convention, which facilitates the searching and retrieval of relevant data. Identical sequences and their annotations are merged together to provide a low-level of redundancy. The SwissProt database places a special emphasis on protein sequences from *Homo sapiens* and model organisms with the goal of creating a representative collection of sequences; however, any data from specialized study groups, genomic-related publications or other literature highlighting a protein's function is also included in SwissProt. The current release of SwissProt (release 48.0, 13 Sep 2005) contains 194,317 records.

With the onset of improved sequencing technology and especially high-throughput genomic sequencing, many more protein sequences are being published than can be analyzed experimentally. Thus, the TrEMBL database includes all protein sequences (excluding those sequences already in SwissProt) from the CDS of nucleotide databases with a minimal level of annotation based solely on automated analyses. The current release of TrEMBL (release 31.0, 13 Sep 2005) contains 2,105,517 records (Boeckmann, et al., 2003). Information on downloading SwissProt and TrEMBL may be found at <http://us.expasy.org/sprot/download.html>.

Structural Databases

PDB – Protein Data Bank

The Protein Data Bank (PDB) is the primary, centralized resource for three-dimensional structures of biological macromolecules. Each structure included in the PDB contains the organism name, SwissProt and GenBank identifiers, PubMed identifiers, and enzyme commission numbers. From this information, links are automatically generated to taxonomy information at NCBI, gene ontology terms, structural genomic targets, and when relevant, to the Kyoto Encyclopedia of Genes and Genomes enzyme database. The PDB stresses data integration and linking to other relevant databases. The PDB may be downloaded from <ftp://ftp.rcsb.org/pub/pdb/> in a variety of formats (Deshpande, et al., 2005).

Domain Databases

Pfam – Protein Family Database

Pfam is a database of protein families and domains. Each domain is represented with a multiple alignment and profile hidden Markov model (HMM), and annotated with a text description, references to pertinent literature, and links to external resources. All entries are classified into one of four categories:

- *Family*: a related group of protein sequences,
- *Domain*: a structural unit able to exist independently,
- *Repeat*: a unit that only occurs in two or more copies, and
- *Motif*: a shorter sequence unit not found as part of a globular domain.

Pfam is comprised of Pfam-A, a manually curated set of domains with detailed information about each family, and Pfam-B, an automatically generated set of domains based on motifs from the PRODOM database (Bru, et al., 2005). The Pfam database is freely available in a variety of formats including HMM's for local searching with the HMMER software (see below), multiple alignments, and the annotation of each family. Because of its high-quality domain families and easy extrapolation of this information onto any given protein sequence, Pfam is an invaluable resource in large-scale functional annotations. In August 2005, the Pfam curators released version 18.0 which contained 7973 families.

The Pfam web server (<http://www.sanger.ac.uk/Software/Pfam/>) provides a number of services for analyzing a protein sequence over the Internet. This includes predicting Pfam domains, signal peptides, transmembrane regions, coiled-coil segments, and low-complexity regions. Complex search options are available such as scanning by taxonomy and looking for particular domain organizations (Bateman, et al., 2004).

SMART – Simple Modular Architecture Research Tool

SMART (Simple Modular Architecture Research Tool) is a database of signaling domains and focuses on characterizing eukaryotic proteins, although numerous prokaryotic domains have also been characterized. Similar to Pfam, each domain is modeled with an HMM and multiple sequence alignment along with descriptive information including publications, functional assignments, and distribution across phyla. Upon signing a licensing agreement, the SMART models may be downloaded for local searching using the HMMER software (see below). The SMART website (<http://smart.embl-heidelberg.de/>) provides enables the user to scan protein sequences for

both Pfam and SMART domains and will identify other sequence features such as transmembrane regions. The January 2004 release of SMART contained 685 domain families (Letunic, et al., 2004).

COG – Clusters of Orthologous Groups

The COG database attempts to classify all conserved genes from available complete genomes into clusters of orthologous groups or COGs. Three or more proteins from distantly related lineages are considered orthologs if they are more similar to each other than any other proteins from the same genome. Because orthologs are considered to share the same function, other uncharacterized or hypothetical genes assigned to a COG may be considered to have an identical function. The COG database has been successfully used to annotate new genomes, characterize a vast number of hypothetical conserved genes, and identify phyletic patterns between various taxonomic groupings (Tatusov, et al., 1997).

Tools

Sequence Similarity Search Tools

BLAST – Basic Local Alignment Search Tool

The first evidence for homologous sequences to a given protein or nucleotide sequence usually begins by searching for similar sequences in a sequence database using BLAST or one of its variants. BLAST rapidly and efficiently identifies statistically significant, pairwise alignments of a query sequence to any other sequences in a target database. The BLAST algorithm accomplishes this feat by approximating the optimal alignment rather than performing a time consuming search of the entire sequence space.

This avoids the construction of many insignificant alignments. While this may result in the loss of some sensitivity, the computational runtime is improved by at least a magnitude of order. Thus BLAST trades some sensitivity in order to gain a huge reduction in the computational time.

The BLAST algorithm comprises three major steps: seeding, extension, and evaluation. Seeding the search space involves detecting all the word-hits, or seeds, between two sequences. A word is a contiguous subsequence of some length W . A word-hit occurs when two sequences share a word that scores at least T as defined in a substitution matrix such as BLOSUM62 (Henikoff and Henikoff, 1992) or PAM250 (Dayhoff, et al., 1978). The seeding step is followed by extending each word-hit in both directions until the alignment score falls X units below the maximum score of the alignment generated thus far. The original version of BLAST performed ungapped alignments; however, the latest version computes gapped alignments in the extension step using dynamic programming (Altschul, et al., 1997). Finally, BLAST evaluates which of the alignments generated by extending the seeds are statistically significant. After sorting the alignments by score, the significance of each alignment in terms of its E-value is determined by the Karlin-Altschul equation. Though the statistical mathematics and complexities underlying the Karlin-Altschul equation and related equations are outside the scope of this dissertation, the E-value represents the likelihood that an alignment would occur randomly. It is calculated from the score, size of sequence search space, and a minor constant. A lower E-value corresponds to a more significant alignment. Any alignments that score above a user-specified E-value are termed high-scoring segment

pairs and are presented to the user along with other relevant information including the database search parameters (Altschul, et al., 1990).

PSI-BLAST – Position Specific Iterative BLAST

PSI-BLAST (Position specific iterative BLAST) enables the detection of weakly related homologs via an iterative searching procedure that scores each alignment with a position specific scoring matrix (PSSM). A PSSM is a two-dimensional frequency matrix that captures the distribution of sequence characters found at each position in a multiple alignment. In other words, the PSSM characterizes the likelihood of finding a particular character at a specific position within a group of aligned sequences. The first PSI-BLAST round executes identically to BLAST using one of the supplied substitution matrices (e.g. BLOSUM62) to score the pairwise alignments. Each subsequent round of PSI-BLAST is preceded by the construction of a multiple alignment from any significant hits, and a corresponding PSSM is used in place of the query sequence to search and estimate the significance of any new hits. Searching continues until convergence (i.e. no more homologs identified) or some arbitrary threshold (e.g. maximum number of iterations) is exceeded. Because new sequences are evaluated based not on a standard scoring matrix but from sequences comprising the PSSM, caution must be taken when deciding which sequences to include for the next iteration in order to prevent “corrupting” the PSSM and thereby scoring unrelated or random sequences as significant (Altschul, et al., 1997).

BLAST databases

Before a sequence can be compared to other sequences using BLAST or PSI-BLAST, the target sequences to be searched against must be compiled into a BLAST database. NCBI regularly releases several, pre-formatted BLAST databases from

sequence databases such as RefSeq and NR. These may be searched via the NCBI BLAST server (<http://www.ncbi.nlm.nih.gov/BLAST/>) or downloaded for local searches using the BLAST executables. Custom BLAST databases may be created from a FASTA-formatted (Pearson and Lipman, 1988) file using the program formatdb. Both formatdb and the BLAST programs are part of the freely available NCBI-Toolkit (ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/CURRENT/).

Multiple Sequence Alignment and Phylogenetics

The simultaneous alignment of multiple protein (or DNA) sequences reveals how a group of proteins are related in terms of sequence similarity. As such, multiple sequence alignments represent an invaluable source of information for many bioinformatic tasks including characterizing protein families, identifying functionally conserved residues, searching for remote homologs, and performing phylogenetic analyses. In this subsection, programs used to generate multiple sequence alignments and phylogenetic trees using various methodologies are presented.

ClustalW

ClustalW is a robust, multiple sequence alignment tool capable of rapidly aligning many nucleotide or protein sequences using a progressive alignment strategy. Initially, all sequences are aligned to each other to produce a distance matrix, which contains the distance from a given sequence to any other sequence in the group based on sequence identity. From this distance matrix, the ClustalW program builds a guide tree, which directs the order in which sequences should be sequentially aligned. Following the guide tree, ClustalW progressively aligns the sequences to each other beginning with the most related sequences and ending with the most divergent sequences. ClustalW has numerous

features and options, such as sequence weighting and adjusting gap penalties.

Furthermore, ClustalW offers the capability to align two distinct multiple alignments to each other (Thompson, et al., 1994).

As with most multiple alignment programs, ClustalW performs well when aligning closely related sequences (greater than 30% identical or with regions of high similarity) and provides mediocre results for more distantly related sequences. The progressive strategy of aligning the most similar sequences first works well in many cases; however, alignment errors at earlier stages of the process are propagated downstream as the remaining sequences are analyzed. Despite these drawbacks, ClustalW executes very rapidly and can align large numbers (thousands) of sequences in a reasonable length of time even on commodity hardware.

T-Coffee

T-Coffee (Tree-based Consistency objective function for alignment evaluation) uses a progressive alignment strategy that takes into account both local and global alignments and information between all pairwise alignments to produce a multiple sequence alignment. The T-Coffee algorithm follows three major stages: 1) constructing a primary library from global and local alignments, 2) extending the primary library, and 3) progressively aligning the sequences using information from these libraries. Initially, global (performed by ClustalW (Thompson, et al., 1994)) and local alignments (generated by LALIGN (Pearson and Lipman, 1988)) are combined into a primary library with each pairwise alignment weighted by its sequence identity. Extending the primary library consists of further weighting each pair of residues by the frequency this pairing occurred in all the other pairwise alignments. Thus, the library contains information about pairwise

alignments between all input pairs rather than information localized to pairing the two most similar sequences. Finally, T-Coffee progressively aligns all the sequences based upon the position-specific weighting scheme of each residue pair from the extended library (Notredame, et al., 2000).

T-Coffee tends to align divergent sequences (less than 30% identical) more accurately than ClustalW. By pairing residues with information from all pairwise alignments, gaps also are more likely to be properly placed. Unfortunately, T-Coffee is heavily time and memory intensive, and it requires substantial hardware in order to align large numbers (more than 200) of sequences in a reasonable period of time.

PCMA

PCMA (profile consistency multiple sequence alignment) multiply aligns sequences using a progressive alignment strategy that balances speed and accuracy. PCMA first aligns all sequences with pairwise identity greater than some threshold (e.g. 40%) using ClustalW. The remaining sequences and groups with lower sequence similarity are aligned using T-Coffee. The authors of PCMA reported accuracy similar to T-Coffee, yet execution time is twenty times faster.

Most multiple alignment tools produce relatively accurate alignments of closely related sequences (i.e. greater than 35% sequence identity), but this accuracy is substantially reduced when attempting to align more divergent sequences. Furthermore, techniques such as the T-Coffee algorithm, which perform reasonably well with divergent sequences, tend to require large amounts of time and memory. By combining the ClustalW and T-Coffee algorithms into a hybrid approach for aligning sequences, PCMA

provides an interface for rapidly aligning sequences without sacrificing the accuracy of T-Coffee (Pei, et al., 2003).

MUSCLE

MUSCLE builds multiple sequence alignments through k -mer distance estimation, progressive alignment using a novel log-odds expectation score, and a tree-dependent restricted partitioning refinement process. In brief, MUSCLE builds a draft progressive alignment based upon a distance matrix created using a k -mer distance measure (Edgar, 2004). The draft alignment is then improved by first recalculating the distance matrix with the Kimura distance measure (Kimura, 1985) and progressively aligning the sequences based on this new model. Finally, MUSCLE refines the alignment using tree-dependent restricted partitioning – an iterative improvement process which divides the alignment based on splitting the tree into two subtrees at a random position and realigning the resulting profiles. This continues until convergence or some user-specified threshold is exceeded. MUSCLE claims to be the fastest and most accurate of ClustalW, T-Coffee, and MAFFT (Kato, et al., 2002) when compared with four, reference alignment test sets (Edgar, 2004). As such, MUSCLE represents one of the best multiple alignment tools available.

SPEM

SPEM constructs protein multiple sequence alignments using profile-profile alignments and predicted secondary structures. For each protein sequence, the secondary structure of each protein is predicted using PSIPRED (Jones, 1999) and a sequence profile derived using PSI-BLAST. SPEM then generates all pairwise alignments using DP, secondary-structure dependent gap penalties, and secondary structure profile

information. From these pairwise alignments a guide tree is built using the neighbor-joining method (Saitou and Nei, 1987) to progressively align all the protein sequences relative to a position specific scoring scheme. SPEM reported an average 7-15% higher accuracy than T-Coffee, MUSCLE, and Probcons (Do, et al., 2005). Due to the amount of time involved in building a PSI-BLAST profile and predicting the secondary structure for each protein, SPEM requires more computational resources than T-Coffee. Thus, SPEM provides relatively high accuracy and is suitable for aligning medium size numbers of divergent sequences (Zhou and Zhou, 2005).

MEGA

MEGA (molecular evolutionary genetics analysis) is a software package to facilitate the investigation of DNA and protein sequence information in an evolutionary context. The latest release, version 3.0, contains numerous capabilities to analyze molecular sequences both manually and automatically including the following: sequence acquisition, alignment, estimating evolutionary distances, and building phylogenetic trees (Kumar, et al., 2004). MEGA was primarily used in this research to produce sequence similarity trees using the neighbor-joining method (Saitou and Nei, 1987).

Consensus

Consensus is a Perl based program for determining the consensus of each position within a protein multiple sequence alignment. Consensus tallies which residues of a column belong to various amino acid groupings (e.g. tiny – alanine, glycine, serine; negative – aspartic acid, glutamic acid; etc.) and displays a character representing that grouping if its representation exceeds a given threshold (e.g. 75%). Such an analysis can reveal functionally conserved sites by identifying various positions whose amino acid

composition suggests a functional role. Consensus is available on the Internet (<http://www.bork.embl-heidelberg.de/Alignment/consensus.html>) or as a Perl script for local execution.

Alignment Shader

Alignment Shader visualizes the conservation within user-specified subgroups of a protein multiple sequence alignment and enables the rapid assessment of conserved residues specific to a particular subgroup. If the BLOSUM consensus for each column of each subgroup exceeds a given threshold (e.g. 75%), Alignment Shader colors the background of the amino acids that belong to the BLOSUM consensus. Alignment Shader is incorporated as part of the bioinformatics platform (Chapter 3) and may be accessed at <http://moscow.biology.gatech.edu/cgi-bin/alignshade.cgi>.

Jalview

Jalview is a Java multiple alignment editor designed for the fast and efficient viewing and editing of large multiple sequence alignments. Jalview has numerous features and capabilities that make it a robust alignment tool including the following: data input and output in a variety of formats, building phylogenetic trees, coloring alignments using an informative color scheme, defining sequence groups, etc. Jalview also supports multiple distinct, integrated views of a sequence alignment (Clamp, et al., 2004).

Secondary Structure Prediction

PSIPRED

PSIPRED predicts the secondary structure of a protein sequence using neural networks and PSI-BLAST profiles. Initially, PSIPRED performs a PSI-BLAST search

with the query protein sequence against a filtered (to remove repetitive regions that would bias the PSI-BLAST search) NR database. PSIPRED then extracts the PSSM from the PSI-BLAST search and feeds this information into a neural network which assigns each residue of the query sequence to one of three secondary structure states: helix, strand, or loop – also known as the Q3 designation (Rost, et al., 1994). At the time of its publication, PSIPRED ranked as a leading secondary structure prediction tool based on its average Q3 score of 76.5 – 78.3% at the 1997 Critical Assessment of methods for protein Structure Prediction experiment (Moult, et al., 1997). Furthermore, PSIPRED remains a leading secondary structure prediction tool based as demonstrated by the evaluation of secondary structure prediction servers (Koh, et al., 2003).

The major principle behind PSIPRED is that sequence variability and conservation information from a group of related proteins enables a much more nearly accurate prediction than just the information from a single protein sequence. Regions of a protein sequence that have a specific function such as an active site or residues for maintaining the structural fold tend to display a high degree of conservation. More variable regions tend to exist on protein surfaces where fewer constraints exist except for a propensity for hydrophilic residues due to the surrounding polar environment. By collecting position specific, amino acid conservation information about related proteins using PSI-BLAST, PSIPRED assesses the solvent accessibility of each amino acid and predicts the secondary structure state based upon the solvent accessibilities typical for secondary structures (Jones, 1999).

Phobius

Phobius may be used to simultaneously predict signal peptides, transmembrane regions, and their topology from a protein sequence. Numerous other programs exist that separately predict transmembrane regions or signal peptides; however, since the patterns for these two secondary elements are highly similar they are prone to false classifications. Phobius' major strength lies in its ability to distinguish between transmembrane regions and signal peptides and to predict both elements in a single step. Furthermore, Phobius is taxonomy-agnostic, meaning that it was trained without respect to a specific kingdom. This enables the blind searching for signal peptides or transmembrane regions without having to know the taxonomy of a given protein sequence *a priori*.

On average, Phobius predicts signal peptides and transmembrane regions with the following accuracies: proteins containing both signal peptides and transmembrane regions, 90%; proteins with transmembrane regions only, 60%; proteins with signal peptides only, 96%; and proteins with neither signal peptides or transmembrane regions, 98%. (Kall, et al., 2004). When compared to SignalP (Bendtsen, et al., 2004) and TMHMM (Krogh, et al., 2001), Phobius predicted significantly fewer misclassifications although it was also slightly less sensitive. Phobius may be freely installed on Unix-based machines to academic institutions, and queries may be submitted online at <http://phobius.cgb.ki.se/>.

Coils

The Coils program identifies coiled-coil regions from a protein sequence. Coiled-coil regions typically consist of two to five helical bundles that twist around each other to form a supercoil and display a characteristic 'knobs-in-holes' packing of the peptide sidechains at the helix-helix interface. Proteins with coiled-coils demonstrate a heptad

repeat where the first and fourth residues (positions a and d) of every seven amino acids are hydrophobic while the other five residues (positions b,c,e,f, and g) are hydrophilic. Heptad repeats readily enable the prediction of coiled-coil regions and are often found in long filamentous proteins such as myosin or chemoreceptors (Lupas, 1997). The Coils program may be downloaded from <ftp://ftp.ebi.ac.uk/pub/software/unix/coils-2.2/>.

Seg

Seg filters out repetitive or regions with low-complexity from protein sequences. Many proteins contain short repeats, homopolymers, or significant subsequences consisting of just a few amino acids types. These uninformative segments can be quite problematic for many bioinformatic analyses such as sequence similarity searches (Wootton and Federhen, 1993). Seg is freely available from <ftp://ftp.ncbi.nih.gov/pub/seg/>.

Domain Architecture Prediction

HMMER

HMMER is a software package for manipulating and working with profile HMMs to analyze protein sequences. Profile HMMs are statistical models that have been successfully used to model protein domains (see introduction). The HMMER software package contains several tools for handling HMM-related tasks including the following: building HMMs from a multiple sequence alignment (hmmbuild), calibrating an HMM (hmmcalibrate), constructing domain libraries, searching sequence databases for domain homologs (hmmsearch, hmmpfam), and aligning sequences to an HMM (hmmalign) (Eddy, 1998). HMMER is an invaluable tool for determining the domain architecture of a

given protein. It has been implemented as a search tool at both the Pfam and SMART web servers and may be installed locally on Unix-based architectures for use with any HMM compatible database (e.g. Pfam, SMART, or customized HMM collections). HMMER is available at <http://hmmer.wustl.edu>.

Visualization

PyMOL, VMD

PyMOL and VMD (Visual Molecular Dynamics) are freely distributed, visualization programs for manipulating, analyzing, and visualizing three-dimensional biological structures. They both support a broad range of functions including multiple simultaneous views of one or more structures, powerful drawing routines of structural features at multiple levels of detail, scripting, measuring distances, and animations. Structures may be loaded from a variety of formats including the popular PDB format from the local hard disk or automatically from the online PDB database. More information about PyMOL including installation instructions is available from the PyMOL website, <http://pymol.sourceforge.net/>. VMD may be downloaded from <http://www.ks.uiuc.edu/Research/vmd/> and is described in Humphrey, et al., 1996.

Weblogo

Weblogo creates sequence logos from a multiple sequence alignment. Sequence logos graphically represent the amino acid conservation at each position within a multiple alignment. The column height indicates the total conservation at this position and the height of each letter within the column denotes that amino acid's conservation (Crooks, et al., 2004).

Blender

Blender is a powerful 3D animation package that was used for creating complex 3D bar graphs and movies (<http://www.blender.org>).

CHAPTER 3

BIOINFORMATICS PLATFORM AND THE MIST DATABASE

Introduction

A thorough, computationally based analysis of microbial signal transduction necessitates the construction of a flexible, integrated, and unified platform for executing programs, storing and retrieving data, and investigating results. Currently, the bioinformatics field is saturated with numerous methodologies, data formats, heterogeneous databases, and incongruent tools. Such fragmentation impedes progress in comparative genomics where integration and comparisons are essential. To address this need for integrated and uniform data management, we deployed a complex assembly of bioinformatic tools, relational database, and knowledge environment. This was implemented using high quality hardware including a powerful dual-Xeon server and a 16-processor Linux cluster. Such an approach requires substantial effort to initially configure, yet it yields a consistent and reliable system that is insulated from external changes. In this chapter, we describe the hardware and software underlying our bioinformatics platform, the central Microbial Signal Transduction (MiST) database, the high-throughput process for identifying signal transduction systems, and the exploratory interface to this platform and MiST.

Hardware and Software

The bioinformatics platform was implemented on state-of-the-art hardware and technology, including a primary server and a 16-processor Linux cluster. The primary

server consisted of dual Intel-based 3.0 gigahertz (Ghz) Xeon processors, four gigabytes (GB) of random access memory (RAM), four SCSI hard drives with 584 GB of storage space that was configured in a redundant array of independent disks (RAID level ten for optimal performance and stability) and the Gentoo Linux operating system (OS). Gentoo (<http://www.gentoo.org>) is a source-based distribution of the Linux OS that provides maximal performance and customization. Connected to this primary server is an 8-node, 16-processor Linux cluster with a peak performance of approximately twenty GigaFLOPs (Floating-point Operations) and 120 GB of NFS-mirrored, storage space. Each node consists of dual Intel-based 2.2 Ghz Xeon processors, two GB of RAM, 120 GB of local storage, and the RedHat Linux OS (<http://www.redhat.com>).

In addition to the numerous open-source packages that come as part of the Gentoo and RedHat Linux distributions, the following software packages also comprised the bioinformatics platform:

- Relational Database Management System (RDBMS): MySQL (<http://www.mysql.com>), PostgreSQL (<http://www.postgresql.org>)
- Programming languages: Perl, Python, C/C++
- Support libraries: eXtensible Markup Language (XML), Common Gateway Interface (CGI), Library for the World-wide-web for Perl (LWP), Database Interface (DBI), Twisted (<http://www.twistedmatrix.com>)
- Web server: Apache 2.0 (<http://www.apache.org>)
- Cluster management: ROCKS (<http://www.rockclusters.org/Rocks/>), Sun Grid Engine (SGE) (<http://gridengine.sunsource.net>)

The MySQL and PostgreSQL RDBMS were chosen for their substantial implementation of the Structured Query Language (SQL) specification, ease of deployment and integration with other tools (e.g. the Perl language, phpMyAdmin, and phpPgAdmin), functionally rich feature sets, high performance, portability, minimal overhead and maintenance, and stability. One of the most popular RDBMS for biological databases, MySQL, is known for its incredible querying speed, web integration, and straightforward configuration; however, due to MySQL's emphasis on speed, it lacks many features associated with other database systems. PostgreSQL constitutes a somewhat slower but more feature-rich RDBMS with support for foreign key references, triggers, views, and nested subqueries. Thus, MySQL is better suited for web applications where speed is critical, and PostgreSQL is preferred for structuring and querying complex data sets.

Perl is the preferred language for bioinformatics because of its scripting capabilities and superior text parsing. Perl also contains useful add-in modules for simplifying the tasks of database management (DBI) and web interfacing (CGI, LWP). For example, the DBI module reduces the complexity of interacting with MySQL or PostgreSQL to a few lines of code and the LWP module enables one to interact with websites from a single command. The Python programming language combines both the powerful simplicity of an interpreted language with the object-oriented capabilities often found in lower-level languages such as C++ or Java. Python in conjunction with the Twisted network library was primarily used as an intermediate networking layer for handling communications between the server and cluster. C/C++ were used to improve performance when the speed of the Perl and Python scripting languages became a

limiting factor; however, C/C++ are quite complex and require careful programming to avoid programming errors such as memory leaks and buffer overflows.

The Apache software package was used to deliver HTML (HyperText Markup Language) content and served as a portal for users to interact with the bioinformatics platform from the Internet. Additionally, customized CGI scripts were developed to execute computationally demanding tasks such as PSI-BLAST and HMMER on the Linux cluster. Scheduling, distributing, and execution of job requests on the cluster were handled using SGE. The cluster is configured and maintained with the Rocks Cluster software.

Whenever possible, data exchange between various programs and scripts was encoded in the XML format to facilitate parsing and to standardize input and output operations. This was particularly important when importing data originating from heterogeneous, external sources in various formats. Non-XML data must first be parsed and then formatted into XML adhering to a specific Document Type Definition (DTD). The DTD describes the components, structure, and organization of a particular set of data in a standard format by associating various XML tags with specific biological information.

Building onto this core infrastructure, several additional bioinformatic databases and tools were installed. These included: NR, RefSeq, SwissProt, Pfam, and SMART, BLAST, PSI-BLAST, ClustalW, T-Coffee, PCMA, MUSCLE, Consensus, PSIPRED, Phobius, Coils, Seg, HMMER, PyMOL, VMD, Weblogo, and Blender. Detailed information regarding these databases and tools may be found in chapter 2, “General Materials and Methods.”

In addition to the open-source software and common bioinformatics programs described above, we developed several utility scripts for performing tedious but useful bioinformatic tasks:

- `aln2fasta` – converts a ClustalW alignment into FASTA format
- `alncdel` – removes gaps from a ClustalW alignment
- `alnidred` – replaces the identifier lines of a ClustalW alignment with unique identifiers that are associated with the original identifier
- `alnlong2short` – transforms a ClustalW alignment from a single block of sequences into multiple blocks of sequences
- `alnshort2long` – transforms a ClustalW alignment from multiple blocks of sequences into a single block of sequences
- `alnstripgaps` – removes any columns that contain only gaps from a ClustalW alignment
- `buildRefSeqFasta` – downloads the complete RefSeq database in FASTA format
- `crc` – calculates the 64-bit cyclic redundancy check hash of FASTA formatted sequences
- `fa2smart` – automatically submits FASTA formatted sequences to the SMART website in order to predict their domain architecture
- `faadups` – removes any duplicate sequences from a FASTA formatted file
- `facount` – counts the number of sequences in a FASTA formatted file
- `faidexp` – replaces the unique identifiers of sequences in a FASTA formatted file with their original identifiers

- `faidred` - replaces the identifier lines of sequences in a FASTA formatted file with unique identifiers that are associated with the original identifier
- `fasplit` – splits a FASTA formatted file into either a user-specified number of files or number of sequences per file
- `fixnr` – concatenates identical sequences from the NR into a single entry
- `gilst2fasta` – downloads from NCBI the FASTA formatted sequence(s) for a list of GI numbers
- `gilst2gbk` – downloads from NCBI the GenBank record(s) for a list of GI numbers
- `gbcount` – counts the number of sequences in a GenBank formatted file
- `hmmcount` – counts the number of HMMs in an HMM database
- `pmatch` – scans a given file for a Perl-compatible regular expression and print any matches
- `taxfromgi` – prints the taxonomy for a list of GI numbers

The MiST database

The vast quantity of genomic data produced by sequencing and annotation projects necessitates the development of databases to organize, comprehend, and analyze this information effectively (Karp, 2001). Thus in order to perform a comprehensive comparative genomic analysis of microbial signal transduction, we designed and implemented the Microbial Signal Transduction (MiST) database using the PostgreSQL (version 8.0) RDBMS. MiST functions as the backbone of the bioinformatic platform by serving as a central, integrated repository of both primary and derived information. The structure of MiST is conceptually divided into four extensible sections (Figure 3.1):

- *Primary genomic data*: DNA and protein sequences and annotation information

- *Derived data*: predicted domain architectures and other secondary features (e.g. transmembrane and low-complexity regions) for each protein, and domain model information
- *Results*: signal transduction domains, classified signal transduction proteins, and statistical information
- *Management*: bookkeeping and tracking of genomic data mining

The primary genomic data section is composed of the Datasource, Organisms, Components, Proteins, and ProteinsData tables. Records within the Components table represent chromosomes, plasmids, or sequence contigs (for draft genomes). Associated with each component are related genes and proteins. Derived data is meant to store and organize any information derived from primary data by computational methods. Each protein's predicted domain architecture, transmembrane regions, signal peptides, coiled-coil regions, and low-complexity regions are stored in the Regions table.

Any proteins determined to be involved in signal transduction are stored in the Results section, which is comprised of the SignalProteins and SignalFamilies tables. Finally, meta-data regarding the status of information flow for each component is maintained in the Tasks table of the management section. The Tasks table identifies any component that has been downloaded to the local file system, parsed and inserted into the database, and tracks whether the domain architecture and various other secondary features have been predicted for the proteins related to each component. These database values are updated with the execution and completion of each task. This infrastructure facilitates the management of both storage and external processing of this data without restricting future development. Incorporating additional tasks simply involves creating

the supporting tables, relating the new tables to existing tables, and adding another column to the tasks table. Perl modules have been developed which interact with the database to enable the rapid construction of custom data files for exporting purposes or processing by other programs.

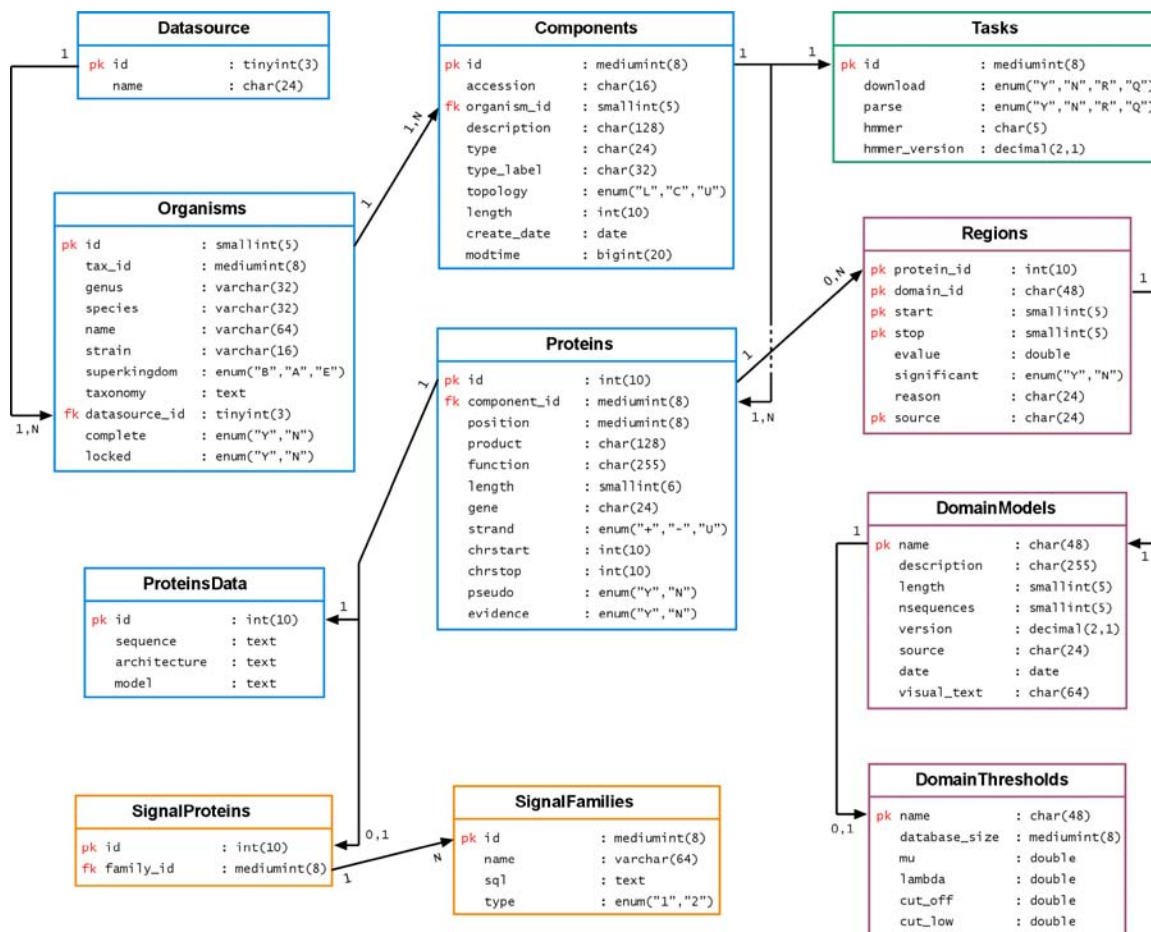


Figure 3.1 Structure of the Microbial Signal Transduction database, MiST. Table names are in bold type and lines indicate the various relationships between tables. Values above the lines signify the cardinality constraints for each relationship. The labels ‘pk’ and ‘fk’ refer to primary key and foreign key, respectively. In some cases, a single column is both a primary and foreign key, in which case only ‘pk’ is listed. The color of each table denotes a particular database section: light blue – primary genomic data; purple – derived data; orange – results; and green – management.

High-throughput identification of signal transduction proteins

The high-throughput identification of signal transduction proteins in prokaryotes involves three major steps (Figure 3.2):

1. All available genomic data is downloaded to the local file system, parsed, and inserted into MiST.
2. The domain architecture and various secondary features of each protein is computationally predicted and inserted into the database.
3. Finally, the set of signal transduction proteins is determined by querying each protein in the database for the presence of conserved signaling domains.

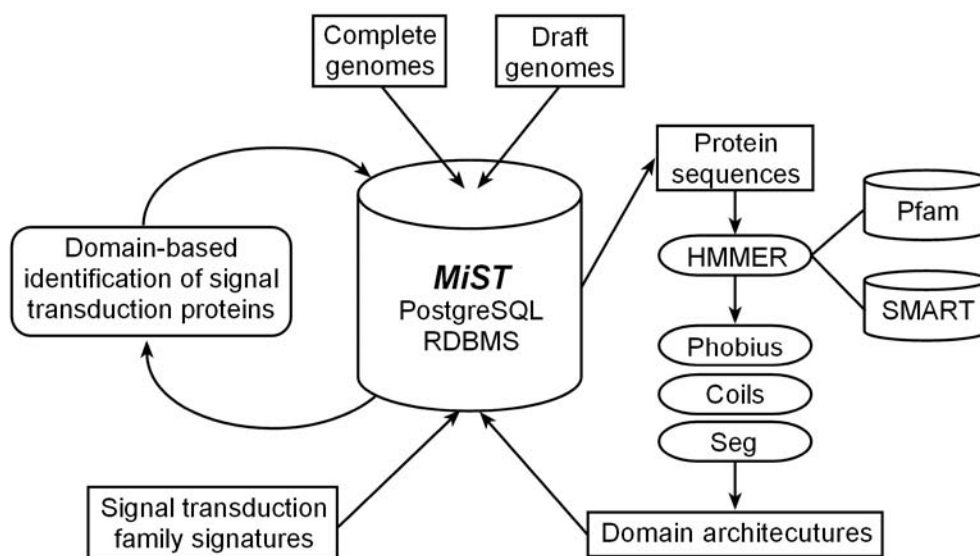


Figure 3.2 High-throughput identification of signal transduction proteins.

Step 1: Retrieval of genomic data and incorporation into MiST

All available complete and draft genomes are downloaded and processed locally so that subsequent information processing is minimally dependent on external sources. Draft genomes consist of several discontinuous shotgun sequences, which may contain overlaps and other sequencing errors. Despite the preliminary nature of this data, many experimental scientists are interested in receiving information about signal transduction proteins from these genomes. Moreover, preliminary analysis shows the largest and most diverse groups of signal transduction proteins are found in several of these draft microbial genomes (e.g. *Magnetospirillum magnetotacticum*). For that reason we have developed a computational infrastructure capable of handling both complete and incomplete genomic data.

Perl scripts take a “snapshot” of all available genomes from the NCBI web and ftp (file transfer protocol) sites, then store this information in XML for further processing. For each genome not present in MiST, its XML encoded genomic data is downloaded to a standard location on the local file system and meta-data regarding this genome’s status recorded in the database. In addition to the nucleotide data and translated protein set, these XML files contain the full RefSeq annotation associated with each gene and protein. Using the XML::Parser module, Perl programs parse this XML data and insert this information into the appropriate tables within MiST.

Because Pfam is openly available and freely provides annotations with the domain models, this library was integrated by recording the model description and parameters in the Pfam tables of MiST. Similarly, any novel domain models may be added using this same process.

Step 2: Prediction and storage of each protein's domain architecture and secondary features

The algorithm for determining the domain architecture of all proteins within MiST is as follows:

1. Create a FASTA file containing all the proteins for those components that have not been hammered.
2. Using the hmmpfam program, search the component FASTA file against Pfam and SMART.
3. Evaluate the significance of each domain prediction and remove insignificant hits.
4. Insert the domain predictions into the Derived data section of MiST.
5. Update the Proteins table to reflect any predicted domains and construct a character string representation of the domain architecture for each protein.

Custom Perl scripts handle the majority of these tasks by “gluing” together the various stages as well as handling all database requests using the DBI module. Due to the large number of protein sequences (roughly one million) and domain models (roughly 7500) to be searched, step three is executed on our local Linux cluster using an “embarrassingly parallel” approach. This involves each slave node searching a proportional slice of the sequence space using a local copy of the HMMER software package and the Pfam and SMART domain libraries. On moderate hardware, searching a single sequence for Pfam domains requires approximately four seconds. On a single desktop server, searching all sequences against all domain models would require more than six months of continuous computation; however, we are able to accomplish this goal in just under a month with our Linux cluster. Each domain is assessed for significance using the respective scoring

scheme from each domain database. Throughout this processing pipeline, the results are exchanged in XML before being copied into MiST. The Proteins table is then updated by assigning to each protein a non-redundant list of domains it contains and a corresponding string representation of its domain architecture. This non-redundant domain list optimizes searching for proteins with specific domains. The string architecture consists of listing each domain prediction in order of its occurrence in the protein sequence and using a tilde (“~”) character for each amino acid that occurs outside a domain prediction. By virtue of the string design, complicated queries may be constructed which include constraints such as a specific sequential ordering of domains and limits regarding the length of amino acids between domains. For example, to select all proteins with an N-terminal REC domain and an unknown C-terminal domain of at least 80 amino acids one could use the following regular expression: “ $^{\wedge}\sim\{0,70\}<\text{REC}>.\sim\{80,\}\$$ ”. Signal peptides, transmembrane regions, coiled-coils, and segments of low-complexity are predicted in a similar fashion, except these predictions are done locally on the server rather than on the cluster.

Step 3: Identifying and classifying signal transduction proteins

Identifying signal transduction proteins entails querying each protein’s domain architecture for the presence of conserved signaling domains. We designated 118 Pfam and SMART domains as signaling domains based on the following: a) their function, b) association with other domains, c) COGs, and d) experience working with signal transduction proteins (Table 3.1). These domains were further classified as input, output, transmitter, or receiver domains and inserted into MiST. Each protein is searched for these domains and inserted into the SignalProteins table if they contain an output,

transmitter, or receiver domain. Input domains often are found in pathways other than signal transduction (e.g. metabolic pathways), and therefore proteins identified solely from an input domain are not classified as belonging to signal transduction. During this process, the Perl script also filters out various domain combinations that indicate a role other than signal transduction. The Perl scripts that constitute this pipeline may be executed on a regular schedule such that new genomes are seamlessly and automatically integrated into the database, and signal transduction blueprints for these organisms are automatically generated.

Table 3.1 Pfam and SMART domains used for identifying signal transduction proteins.

Domain	Source	Classification	Function
<i>BLUF</i>	Pfam	Input	Cofactor binding
<i>FeS</i>	Pfam	Input	Cofactor binding
<i>Fer4</i>	Pfam	Input	Cofactor binding
<i>Hemerythrin</i>	Pfam	Input	Cofactor binding
<i>HhH-GPD</i>	Pfam	Input	Cofactor binding
<i>NIR_SIR</i>	Pfam	Input	Cofactor binding
<i>NIR_SIR_ferr</i>	Pfam	Input	Cofactor binding
<i>Nitro_FeMo-Co</i>	Pfam	Input	Cofactor binding
<i>Phytochrome</i>	Pfam	Input	Cofactor binding
<i>Aminotran_1_2</i>	Pfam	Input	Enzymatic
<i>Arch_ATPase</i>	Pfam	Input	Enzymatic
<i>Citrate_synt</i>	Pfam	Input	Enzymatic
<i>Cyanate_lyase</i>	Pfam	Input	Enzymatic
<i>EPSP_synthase</i>	Pfam	Input	Enzymatic
<i>FmdA_AmdA</i>	Pfam	Input	Enzymatic
<i>GATase_2</i>	Pfam	Input	Enzymatic
<i>Glucokinase</i>	Pfam	Input	Enzymatic
<i>Glycos_trans_3N</i>	Pfam	Input	Enzymatic
<i>Glyoxalase</i>	Pfam	Input	Enzymatic
<i>HEAT_PBS</i>	Pfam	Input	Enzymatic
<i>HEM4</i>	Pfam	Input	Enzymatic
<i>NTP_transf_2</i>	Pfam	Input	Enzymatic
<i>NUDIX</i>	Pfam	Input	Enzymatic
<i>Nitroreductase</i>	Pfam	Input	Enzymatic
<i>PALP</i>	Pfam	Input	Enzymatic
<i>PTS-HPr</i>	Pfam	Input	Enzymatic
<i>PTS_EIIC</i>	Pfam	Input	Enzymatic
<i>Peptidase_M23</i>	Pfam	Input	Enzymatic
<i>PfkB</i>	Pfam	Input	Enzymatic
<i>Pribosyltran</i>	Pfam	Input	Enzymatic
<i>Pyr_redox</i>	Pfam	Input	Enzymatic
<i>Rhodanese</i>	Pfam	Input	Enzymatic
<i>SKI</i>	Pfam	Input	Enzymatic
<i>peroxidase</i>	Pfam	Input	Enzymatic
<i>CBS</i>	Pfam	Input	Protein-protein interaction
<i>HAMP</i>	Pfam	Input	Protein-protein interaction
<i>TPR_1</i>	Pfam	Input	Protein-protein interaction
<i>TPR_2</i>	Pfam	Input	Protein-protein interaction
<i>TPR_3</i>	Pfam	Input	Protein-protein interaction
<i>TPR_4</i>	Pfam	Input	Protein-protein interaction
<i>ACT</i>	Pfam	Input	Small-molecule binding
<i>Ada_Zn_binding</i>	Pfam	Input	Small-molecule binding
<i>AlkA_N</i>	Pfam	Input	Small-molecule binding
<i>AraC_binding</i>	Pfam	Input	Small-molecule binding
<i>Autoind_bind</i>	Pfam	Input	Small-molecule binding
<i>CHASE</i>	Pfam	Input	Small-molecule binding
<i>Cache</i>	Pfam	Input	Small-molecule binding
<i>Diacid_rec</i>	Pfam	Input	Small-molecule binding
<i>FHA</i>	Pfam	Input	Small-molecule binding
<i>Fe_dep_repr_C</i>	Pfam	Input	Small-molecule binding
<i>FeoA</i>	Pfam	Input	Small-molecule binding
<i>GAF</i>	Pfam	Input	Small-molecule binding
<i>HMA</i>	Pfam	Input	Small-molecule binding
<i>LysR_substrate</i>	Pfam	Input	Small-molecule binding
<i>PAS</i>	Pfam	Input	Small-molecule binding
<i>PAS</i>	SMART	Input	Small-molecule binding
<i>PAC</i>	SMART	Input	Small-molecule binding
<i>PBP</i>	Pfam	Input	Small-molecule binding
<i>PBP</i>	Pfam	Input	Small-molecule binding

Table 3.1 (continued)

<i>PPP</i>	Pfam	Input	Small-molecule binding
<i>SIS</i>	Pfam	Input	Small-molecule binding
<i>STAS</i>	Pfam	Input	Small-molecule binding
<i>TOBE</i>	Pfam	Input	Small-molecule binding
<i>TetR_C</i>	Pfam	Input	Small-molecule binding
<i>V4R</i>	Pfam	Input	Small-molecule binding
<i>cNMP_binding</i>	Pfam	Input	Small-molecule binding
<i>CHASE2</i>	Pfam	Input	Unknown function
<i>CHASE3</i>	Pfam	Input	Unknown function
<i>CHASE4</i>	Pfam	Input	Unknown function
<i>MASE1</i>	Pfam	Input	Unknown function
<i>MASE2</i>	Pfam	Input	Unknown function
<i>MHYT</i>	Pfam	Input	Unknown function
<i>TrkA_C</i>	Pfam	Input	Unknown function
<i>Arc</i>	Pfam	Output	DNA-binding
<i>Arg_repressor</i>	Pfam	Output	DNA-binding
<i>AsnC_trans_reg</i>	Pfam	Output	DNA-binding
<i>Crp</i>	Pfam	Output	DNA-binding
<i>CtsR</i>	Pfam	Output	DNA-binding
<i>DeoR</i>	Pfam	Output	DNA-binding
<i>Fe_dep_repress</i>	Pfam	Output	DNA-binding
<i>GerE</i>	Pfam	Output	DNA-binding
<i>GntR</i>	Pfam	Output	DNA-binding
<i>HTH_1</i>	Pfam	Output	DNA-binding
<i>HTH_10</i>	Pfam	Output	DNA-binding
<i>HTH_3</i>	Pfam	Output	DNA-binding
<i>HTH_5</i>	Pfam	Output	DNA-binding
<i>HTH_6</i>	Pfam	Output	DNA-binding
<i>HTH_7</i>	Pfam	Output	DNA-binding
<i>HTH_8</i>	Pfam	Output	DNA-binding
<i>HTH_AraC</i>	Pfam	Output	DNA-binding
<i>IclR</i>	Pfam	Output	DNA-binding
<i>LacI</i>	Pfam	Output	DNA-binding
<i>LytTR</i>	Pfam	Output	DNA-binding
<i>MarR</i>	Pfam	Output	DNA-binding
<i>MerR</i>	Pfam	Output	DNA-binding
<i>PadR</i>	Pfam	Output	DNA-binding
<i>RHH_1</i>	Pfam	Output	DNA-binding
<i>ROS_MUCR</i>	Pfam	Output	DNA-binding
<i>TetR_N</i>	Pfam	Output	DNA-binding
<i>Trans_reg_C</i>	Pfam	Output	DNA-binding
<i>EAL</i>	Pfam	Output	Di-guanylate cyclase
<i>GGDEF</i>	Pfam	Output	Di-guanylate cyclase
<i>HD</i>	Pfam	Output	Hydrolase
<i>Guanylate_cyc</i>	Pfam	Output	Other
<i>LytR_cpsA_psr</i>	Pfam	Output	Other
<i>Rrf2</i>	Pfam	Output	Other
<i>RseA_N</i>	Pfam	Output	Other
<i>PP2C_SIG</i>	SMART	Output	Phosphatase
<i>Pkinase</i>	Pfam	Output	Protein kinase
<i>ANTAR</i>	Pfam	Output	RNA-binding
<i>CsrA</i>	Pfam	Output	RNA-binding
<i>Response_reg</i>	Pfam	Receiver	Response regulator
<i>HATPase_c</i>	Pfam	Transmitter	Histidine kinase
<i>HATPase_c</i>	SMART	Transmitter	Histidine kinase
<i>HisKA</i>	Pfam	Transmitter	Histidine kinase
<i>HisKA</i>	SMART	Transmitter	Histidine kinase
<i>MCPsignal</i>	Pfam	Transmitter	MCP
<i>MA</i>	SMART	Transmitter	MCP

Exploratory Knowledge Environment

In order to provide an effective means for exploring and utilizing the bioinformatics platform beyond local data mining, we implemented a web-based knowledge environment. The majority of programs comprising the bioinformatics platform lack graphical user-interfaces and are invoked from the command-line interface of a computer shell. While this approach typically provides the user with the most options and capabilities, the non user-friendly terminal often acts as a barrier to research for biologists – the majority of which are not comfortable or familiar with this mode of operation. To alleviate this problem, we built a user-friendly, web interface that enables users to perform basic bioinformatics related tasks as well as query and interact with the MiST database.

BLAST, HMMER, Consensus, and an alignment-shading program were implemented as part of the knowledge environment. To provide scalability and boost performance, BLAST and HMMER were remotely executed on the Linux cluster via Python scripts using the Twisted network library. The typical path of a BLAST or HMMER job is as follows:

- 1) The user inputs any sequence(s) to be searched via the website,
- 2) CGI scripts encode this request in XML and transmit the job request to the cluster,
- 3) the request is decoded and executed,
- 4) the program output is encoded into XML and transmitted back to the CGI script on the Apache web server, and finally,
- 5) the XML results are appropriately presented to the user.

A CGI wrapper program (Figures 3.3, 3.4) manages PSI-BLAST requests with a number of additional, useful features that include the following: customized XML-encoded input/output to PSI-BLAST, the ability to save and load searches at any iteration, download sequences from selected hits, view the domain architectures for any selected hits, and an improved visualization and user-interface. The front-end to HMMER enables the user to search multiple sequences against either Pfam or SMART and visualizes each protein's domain architecture. Consensus and the alignment shader are executed on the web server as CGI scripts. Screenshots of each of these programs in action are provided in Figures 3.5-3.8.

PSI-BLAST 3

Enter Sequence Data (Raw):
AGFLGVEKIDKASNEILSQEVKIGEYFSRVRANILYMRMYEKDAF ININNPDKIAEYEKKWTEKKGRLEW
LGKLGKLLDDKEKGQYAAIQENYKQYVDGFGQLMGQIKSGAITSTQQANEAMKPVKEAAQAIEKLSTEGN
RAAYEMSDKKTEVDAIGQRS

Reset Search

Database: nr (13 May 2005) Filters:
☒ Low complexity
☐ Mask lookup table only

E-value Threshold: 0.01
Maximum Threshold: 10

Matrix: BLOSUM62 Other Options:
☒ Composition-based statistics
☐ Compute Smith-Waterman Alignments

No. Descriptions: 500
No. Alignments: 500

Load Previous Search: Browse... Load

Figure 3.3 Web interface to PSI-BLAST. Input spaces are provided for the query sequence, database to search, and other various options. Users may also load previous searches using this page.

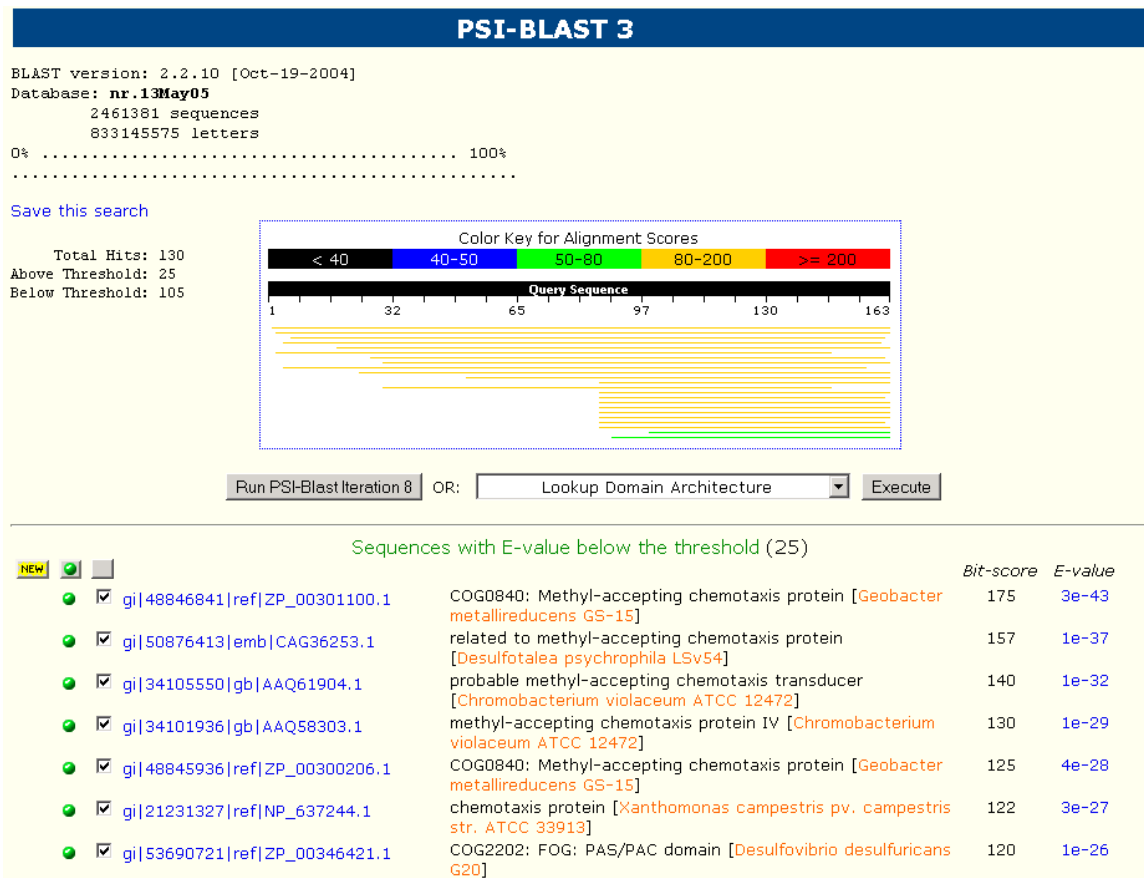


Figure 3.4 Output of results from a web-based PSI-BLAST search. Descriptive information about the search includes the BLAST version, database size, and number of hits. A graphical output displays hits to the query sequence as horizontal lines and the color of each line indicates its score. The search may be continued additional iterations or the results from this search may be used in another analysis (e.g. view the domain architectures for the selected sequences). Significant BLAST hits and information about each pairwise alignment are displayed below the graphical overview.

Consensus

Upload Clustal File:

OR Input clustal alignment data below:

Bcep_46317933

--TQQEFDFPDDVTLMSSTDADSIITYANTTFAYVSGFSTDELVGQPHNVRHPDMPKEAFADNMWATLKRRC

Bcep_46322894

PVTQQEFDFPDDVTLMSSTDADSIITYANTTFAYVSGFSTDELVGQPHNVRHPDMPKEAFADNMWATLKRNC

Bmal_52423246

-----MSTTDPHGRITYANATFVHVSGFSSDEIVGAPHNVVRHPDMPRDAFADNMWATLKRRC

Bpse_52211725

--TQHEFELPDDATLMSTTDPHGRITYANATFVHVSGFSSDEIVGAPHNVVRHLDMPRDAFADNMWATLKRRC

Bmal_52428136

--TQREYDFPDDATLMSTTDQSYVITYANAAFIVQVSGFERDEIIGEPHNVRHPDMPTEAFADNMWATLKAC

Ecol_26110084

--TQONTPLADDTLMSTTDLSQYITHANDTFVQVSGYTLQELQCGQPHNVRHPDMPKAAAFADNMWFLTKKC

Ecol_13363427

--TQONTPLADDTLMSTTDLSQYITHANDTFVQVSGYTLQELQCGQPHNVRHPDMPKAAAFADNMWFLTKKC

Consensus Thresholds:

☒ 50
 ☐ 55
 ☒ 60
 ☐ 65
 ☒ 70
 ☐ 75
 ☒ 80
 ☐ 85
 ☒ 90
 ☐ 95

Figure 3.5 Web interface to the Consensus tool. Users upload ClustalW alignments or input the alignment directly into the provided text area. Checkboxes specify the consensus levels to produce.

Consensus

Please wait... Complete.

consensus/90%

..st..h.hspt..lhopTD.pu.lt.ANtsFnpSuep.pEh.GtPHNhVRHPDMP.tAFADMW.sL+tcGcPwulVKMRKsGsaYVWhAMh.PhhctGp..GY.SlRstssc.

consensus/80%

..Tptphphs-stslhSTTDhpuhIsaANtsFnp1SGas.-E1.GpPHNhVRHPDMPttAFADMWsTL+tcGcPwulVKMRKsGDeYVW+ANAsPlh+pGp..GyhSVRTpsst

consensus/70%

..TppEhchP-ssLMSTTDhpuhIoaANssFnpVSGFo.-E1.GpPHNhVRHPDMPtcAFADMWuTLKtGcPwulVKMRKsGDHYVVRANAsPvHsGpstGYMSVRT+ss+t

consensus/60%

..TQpEachPDDsTLMSTTDspShIITYANsAF1QVSGFop-E11GQPHN1VRHPDMP+-AFADMWATLKsGEPwTALVKMRKNGDHYVVRANAlPvHsGpsGYMSVRT+soR-

consensus/50%

..TQpEa-hPDDsTLMSTTDspShIITYANsAF1QVSGFSpEE11GQPHN1VRHPDMP+EAFAADMWATLKuGEPwTALVKMRKNGDHYVVRANAlPvHsGQssGYMSVRTKpOR-

Daro_46140309

--TDVETRLPEGQFIYSRTDLKGVITEANEAFAKISAYFREEMIGPHNMVRHPDMPAAAFADMWNDLRACRPwRGVVKMRFRDGGYVWLANASPIREHGQIVGYQSVRTAPGR-

Daro_53729639

PVINVETHLPEGEFIYSSTDLOGNLVEANEAFAKISNFREEMIGPHNMVRHPDMPAAAFADMWNDLRACRPwRGVVKMRFRDGGYVWLANASPIREHGQIVGYQSVRTAPGR-

Reut_53762384

--TDNEYRLPSDEVIITKTDAGNIEYANQAFRRSSGYDRAEIIIGQPHN1VRHPDMPAAAFADMWATIRGGTPWTGCVVKMRKDG6GYVWLANITPVFDGKPSGYLSVRTAPTKA

Rgel_47573623

--AAGIVSAVQQAFLITTTDLQGAISFANKAFLQTCGYAMEQVLGAPHSIVRHPDMPKVFADMWAVLHAGRPTGLVKMRSSSDGAFVVKANIIIPMKDRQTVGFTSVQCFFADA

Bcep_46322895

--TQREFEFPDDATLMSTTDANSYIYANAAFIVQVSGFSPEEIEGQPHNVRHPDMPKEAFADMWATLKNGEPwTALVKMRKNGDHYVVRANsVWVVRNGQPTGYMSVRTKASPD

Bcep_46317934

--TQREFDFPDDATLMSTTDANSYIYANAAFIVQVSGFSPEEIEGQPHNVRHPDMPKEAFADMWATLKNGEPwTALVKMRKNGDHYVVRANAlPVMRNGEPKGYMSVRTKATPD

Bfun_48783787

--TQQEFEPDDATLMSTTDTSYVITYANAAFIVQVSGFSLEEIEGQPHNVRHPDMPKEAFADMWATLKNGEPwSALVKMRKNGDHYVVRANATPVVRNGQPAAGYMSVRTQASRE

Rso1_17431698

--TQREYEFDDATLMSTTDTSYIAYANAAFVQVSGFSREEIEGQPHN1VRHPDMPPEAFADMWATLKNGEPwSALVKMRKNGDHYVVRANATPVVRNGRPAAGYMSVRTKPTPD

Bcep_46317933

--TQQEFDFPDDVTLMSSTDADSIITYANTTFAYVSGFSTDELVGQPHNVRHPDMPKEAFADNMWATLRRGEPwTALVKMRKNGDHYVVRANAlPVMRNGEPQGYMSVRTKAPHD

Bcep_46322894

PVTQQEFDFPDDVTLMSSTDADSIITYANTTFAYVSGFSTDELVGQPHNVRHPDMPKEAFADNMWATLKNGEPwTALVKMRKNGDHYVVRANAVPVMRNGAPHGYSVRTKAPRD

Bmal_52423246

-----MSTTDPHGRITYANATFVHVSGFSSDEIVGAPHNVVRHPDMPRDAFADNMWATLKNGEPwTALVKMRKNGDHYVVRANAVPVIRGGQTGYMSVRTKAPARA

Bpse_52211725

--TQHEFELPDDATLMSTTDPHGRITYANATFVHVSGFSSDEIVGAPHNVVRHLDMPRDAFADNMWATLKNGEPwTALVKMRKNGDHYVVRANAVPVIRGGQTGYMSVRTKAPARA

Bmal_52428136

--TQREYDFPDDATLMSTTDQSYVITYANAAFIVQVSGFERDEIIGEPHNVRHPDMPTEAFADNMWATLKAGRSWTAIVKMRKNGDHYVVRANATPVVRNGQLVGYMSVRTKPSRE

Ecol_26110084

--TQONTPLADDTLMSTTDLSQYITHANDTFVQVSGYTLQELQCGQPHNVRHPDMPKAAAFADMWFLTKKGEpWSGIVKMRKNGDHYVVRANAVPVHREGKISGYMSIRTRATDE

Ecol_13363427

--TQONTPLADDTLMSTTDLSQYITHANDTFVQVSGYTLQELQCGQPHNVRHPDMPKAAAFADMWFLTKKGEpWSGIVKMRKNGDHYVVRANAVPVHREGKISGYMSIRTRATDE

consensus/90%

..st..h.hspt..lhopTD.pu.lt.ANtsFnpSuep.pEh.GtPHNhVRHPDMP.tAFADMW.sL+tcGcPwulVKMRKsGsaYVWhAMh.PhhctGp..GY.SlRstssc.

consensus/80%

..Tptphphs-stslhSTTDhpuhIsaANtsFnp1SGas.-E1.GpPHNhVRHPDMPttAFADMWsTL+tcGcPwulVKMRKsGDeYVW+ANAsPlh+pGp..GyhSVRTpsst

consensus/70%

..TppEhchP-ssLMSTTDhpuhIoaANssFnpVSGFo.-E1.GpPHNhVRHPDMPtcAFADMWuTLKtGcPwulVKMRKsGDHYVVRANAsPvHsGpstGYMSVRT+ss+t

consensus/60%

..TQpEachPDDsTLMSTTDspShIITYANsAF1QVSGFop-E11GQPHN1VRHPDMP+-AFADMWATLKsGEPwTALVKMRKNGDHYVVRANAlPvHsGpsGYMSVRT+soR-

consensus/50%

..TQpEa-hPDDsTLMSTTDspShIITYANsAF1QVSGFSpEE11GQPHN1VRHPDMP+EAFAADMWATLKuGEPwTALVKMRKNGDHYVVRANAlPvHsGQssGYMSVRTKpOR-

Figure 3.6 Output of results from the web-based Consensus tool. The alignment is redisplayed with the consensus above and below.

69

Alignment Shader

Upload Clustal File:

OR Input clustal alignment data below:

```

Daro_46140309
--TDVETRLPEGQFIYSRTDLKGVITEANEAFQISAYRREEMLGHNHNVHRPDMPAFAAFADNMNDLRAC
Daro_53729639
PVTNVETHLPEGEFIYSSTDLQGNLVEANEAFKISNFSREEMIGQPHNVHRPDMPAFAAFADNMNDLRAC
Reut_53762384
--TDNEYRLPSDEVIITRTDAQGNIEYANQAFRRSSGYDRAEIIIGQPQNIVRHPDMPAFAAFADNMWATIRGC
Rgel_47573623
--AAGIVSAVQQAFLITTTDLQGAISFANKAFLQTCGYAMEQVLGAPHSIVRHPDMPPKVFADNMWAVLHAC
Bcep_46322895
--TQREFEFPDDATLMSTTDANSYIQYANAAFIQVSGFSPEEIEGQPHNVVRHPDMPEAFADNMWATLKNQ
  
```

Alignment Shading Threshold (%):

☐ 50 ☐ 55 ☐ 60 ☐ 65 ☐ 70 ☒ 75 ☐ 80 ☐ 85 ☐ 90 ☐ 95

Sequence Groups:

Start: Stop: Color:

Start: Stop: Color:

Figure 3.7 Web interface to the Alignment Shader tool. Users upload ClustalW alignments or input the alignment directly into the provided text area. The user then describes the groups of sequences within the alignment, what color to shade them, and at what similarity threshold a column must surpass for it to be shaded.

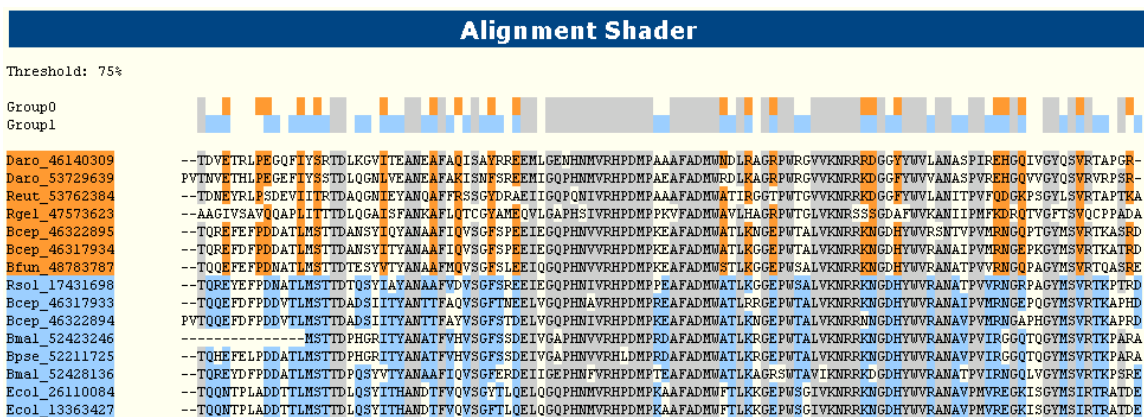


Figure 3.8 Output of results from the web-based Alignment Shader tool.

A sophisticated web interface has also been built for exploring the MiST database. Initially, the user is presented with a list of genomes in MiST organized by an adjustable level of taxonomy (Figure 3.9). Organisms of interest may be selected by clicking the checkboxes beside the organism name and then pushing the ‘Select Organisms’ button. This action presents the analysis selection page (Figure 3.10), which provides the ability to query the selected organisms by domain or domain architecture, description, GI, and MiST identifier. Additionally, the user may view the number of domains in each of the selected organisms by inputting a list of comma-separated domains in the ‘Domain counts’ field and clicking ‘Count’. Clicking on an organism name displays both general and specific information about signal transduction for that organism including: descriptive data about the genome and proteome (e.g. size, chromosomes, nucleotide frequencies, etc.), a graphical representation of its signal transduction profile, querying options, and lists of the one- and two-component systems identified in this microbe (Figure 3.11). The profile presents an overall view of the number of signaling systems separated into functional categories. For example, *E. coli* contains 221 DNA-binding domains, clearly indicating a substantial level of gene regulation (Figure 3.12).

Organisms in MIST (318)

Taxonomic Level: [Kingdom](#) [Phyla](#) [Class](#) [Order](#) [Family](#) [Select Organisms](#)

- ☐ Archaea (25)
 - ☐ Crenarchaeota (5) [Show/Hide](#)
 - ☐ Euryarchaeota (19) [Show/Hide](#)
 - ☐ Nanoarchaeota (1) [Show/Hide](#)
- ☐ Bacteria (293)
 - ☐ Actinobacteria (24) [Show/Hide](#)
 - ☐ Aquificae (1) [Show/Hide](#)
 - ☐ Bacteroidetes (5) [Show/Hide](#)
 - ☐ Chlamydiae (9) [Show/Hide](#)
 - ☒ *Chlamydia muridarum*
 - ☒ *Chlamydia trachomatis*
 - ☐ *Chlamydothrix abortus* S26 3
 - ☒ *Chlamydothrix caviae*
 - ☒ *Chlamydothrix pneumoniae* AR39
 - ☐ *Chlamydothrix pneumoniae* CWL029
 - ☐ *Chlamydothrix pneumoniae* J138
 - ☐ *Chlamydothrix pneumoniae* TW 183
 - ☐ *Parachlamydia* sp. UWE25
 - ☐ Chlorobi (1) [Show/Hide](#)
 - ☐ Chloroflexi (2) [Show/Hide](#)
 - ☐ Cyanobacteria (14) [Show/Hide](#)
 - ☐ Deinococcus-Thermus (3) [Show/Hide](#)
 - ☐ Firmicutes (78) [Show/Hide](#)
 - ☐ Fusobacteria (2) [Show/Hide](#)
 - ☐ Planctomycetes (1) [Show/Hide](#)
 - ☐ Proteobacteria (146) [Show/Hide](#)
 - ☐ Spirochaetes (6) [Show/Hide](#)
 - ☒ *Borrelia burgdorferi*
 - ☐ *Borrelia garinii* PBI
 - ☐ *Leptospira interrogans* serovar Copenhageni
 - ☐ *Leptospira interrogans* serovar Lai
 - ☐ *Treponema denticola* ATCC 35405
 - ☒ *Treponema pallidum*
 - ☐ Thermotogae (1) [Show/Hide](#)

Search for specific protein (Separate multiple ids with spaces):

☒ GI / Other ID ☐ MIST protein ID

• [MIST Statistics](#)

Figure 3.9 Entry web page to the MiST database. Available genomes are listed according to their taxonomy. Clicking the individual checkboxes beside each organism name selects them for further investigation. Users may also search by GI number or MiST identifiers for a specific protein.

Selected Organisms (11):

Bacteria

Chlamydiae

☒ *Chlamydia muridarum*
☒ *Chlamydia trachomatis*
☒ *Chlamydophila caviae*
☒ *Chlamydophila pneumoniae AR39*

Firmicutes

☒ *Bacillus anthracis Ames*
☒ *Bacillus subtilis*
☒ *Geobacillus kaustophilus HTA426*

Proteobacteria

☒ *Escherichia coli K12*
☒ *Pseudomonas aeruginosa*

Spirochaetes

☒ *Borrelia burgdorferi*
☒ *Treponema pallidum*

Select type of search:

Type of search

☒ Domain Search
☐ Advanced Domain Search
☐ Description
☐ GI / Other ID
☐ MIST Protein ID

Search Terms
Domain counts:

Figure 3.10 Analysis selection web page for the MiST database. Each of the previously selected organisms may be searched by domain, description, GI, or MiST identifiers. Clicking on an individual organism name displays various information for that particular organism.

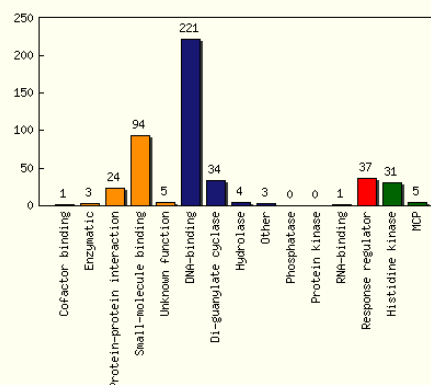
Escherichia coli K12

Genome Summary

- Size: 4.64 MB (4639675 bp)
 - Chromosomes / Plasmids / Contigs: 1
 - Genes: 4410
 - Proteins: 4237
 - Nucleotide frequencies
 - A 24.6% (1142228 bp)
 - C 25.4% (1179554 bp)
 - G 25.4% (1176923 bp)
 - T 24.6% (1140970 bp)
 - N 0.0% (0 bp)
- GC content: 50.8% (2356477 bp)

Signal Transduction Profile

	Proteins	Systems
Two-component:	68	37
One-component:	230	230
Total:	298	267



Select type of search:

- Type of search
- ☒ Domain Search
 - ☐ Advanced Domain Search
 - ☐ Description
 - ☐ GI / Other ID
 - ☐ MIST Protein ID
- Search Terms
-
-

Two component proteins

GI :: Other ID	Description	Inputs	Outputs	Domain Architecture
16128384	positive response regulator for pho regulon, sensor is PhoR (or CreC)	Response_reg	Trans_reg_C	Pfam (Details) SMART (Details)
16128385	positive and negative sensor protein for pho regulon	PAS	HATPase_c	Pfam (Details) SMART (Details)

One component proteins

GI :: Other ID	Description	Inputs	Outputs	Domain Architecture
16128014	transcriptional activator of cation transport (LysR family)		HTH_1	Pfam (Details) SMART (Details)
16128058	transcriptional regulator for ara operon	AraC_binding	HTH_AraC	Pfam (Details) SMART (Details)
49175996	probable transcriptional activator for leuABCD operon	LysR_substrate	HTH_1	Pfam (Details) SMART (Details)
16128073	transcriptional regulator of the control of carbon and energy metabolism (GalR/LacI family)	PBP	LacI	Pfam (Details) SMART (Details)
16128106	transcriptional regulator for pyruvate dehydrogenase complex		GntR	Pfam (Details) SMART (Details)
16128153	deoxyguanosine triphosphate triphosphohydrolase		HD	Pfam (Details) SMART (Details)

Figure 3.11 Organism specific page for the MiST database containing descriptive information about the genome, signal transduction profile, querying options, and lists of one- and two-component proteins found in this organism.

From the organism page, clicking on an individual protein hyperlink such as GI number 16128385 of the *E. coli* organism page, displays an information-rich protein/gene page (Figure 3.12). One may view basic information about this gene and protein including their RefSeq annotations, predicted Pfam and SMART domains, and the genome neighborhood. The amino acid and DNA sequences may be retrieved from the ‘Sequence’ hyperlinks. Additionally, one may extract a user-specified length of upstream or downstream DNA from the gene. The domain architecture section presents a linear visualization of the protein’s domain architecture and other secondary features. Any predicted domains are displayed as white boxes with the domain name printed inside this box. Other secondary features are visualized by filled rectangles with the following colors: signal peptides, red; transmembrane regions, blue; coiled-coils, green; and low-complexity regions, purple. The chromosome view reveals any neighboring genes. The currently selected gene is shown in blue and other surrounding genes are colored in gray. Each gene is clickable and will display a similar page for that gene.

Escherichia coli K12

Protein: 226993 (431 aa) [Sequence](#)

HLRLSUKRLVLELLLCCLPAPILGATFOYLPNLLASVTGLL IUNTUML
LRLSVMLNARSRPTPPGGSHPEPLLYLNHMLPMKXQSPDEL GEL IEDP
PSGAESLPDAVALTTEEGGIFUCNGLAQQLGLRWEDENGQNLMLLRYP
EFTQVLEKTRDTSFPLMLVLTGEHLEIDAMPYTHKQLLVAARDUTQHQL
EGARMTTFARUSHELEKPLTULQGVLEHNEQPLEGAUREKALNTHDEQT
QHEDLVKULLLTSEIEAFTHLLNEKQVOPHMLVUSEANTLSQKQST
TTTEIDWGLDGGREQLSPAIEHLPYDAHNTPEFTHTITDMDQYHGA
EFSVEDNGPGIAPERHPLTERFVVDKARSRQTGGSGGLATVUGHANWH
HEURLNIESTUGKGRTRFSVPIPERLIACNSD

GI: 16128385

NCBI Refseq Annotation

Accession: NP_414934.1

Description: positive and negative sensor protein for pho regulon

Function: N

Experimental: N

COG ID:

Other notes: sensory histidine kinase in two-component regulatory system with PhoB, regulation of Pi uptake, senses Pi; go_component: inner membrane [goid 0019866]; go_process: phosphorus metabolism [goid 0006793]; go_process: protein modification [goid 0006464]

Gene: 234630 (1296 bp) [Sequence](#)

GTGCTGGAAACGGCTGTCTGTGGAAAAGGCTGGTGTCTGGAGCTGCTACTTTG
CTGCTCCCGGCTTTATCTCTGGGTGCATTTTTCGTTAACTGCCTGGT
TTTGTCTGGCATCGGTAAACAGACTGCTATCTGGCATTTCTGGAAATTA
TTGGCGCTTTCATGGTGGCTGTGGGTGGATCGAGTATGACCCGGCCACC
GGGGCTGGTGGTACCTGGGAACCGCTACTATACGGCTTACACAGATGCAGC
TGGAAATAAAAAACGCCCGCTGAACTGGGCAATCTGATTAAACGCTTT
CTGAGCGCGGGAGTCTGCTCCCGGAGCGGGTGGTGTGACACGGAGA
GGGCGGTATTTCTGGGTAACTGCTCTGGCGCAACAATCTTGGTTGCG
GCTGGCGGGAAGATAACGGCGAGAACATCTTAACCTACTGCTTACCGG
GAGTTTACGCAATATCTGAAAACCGGTGATTTTCTCGCCCGCTCAATCT
GGTGTCAACACGGCGGCGCATCTGGAAATTCGGGTCAAGCTTATACCC
ACAAACAGTTGCTGATGGTGGCGCTGATGTCAACGCAANTGCATCAACTG
GAAGGGGGCGGGCGTAACTTTTGGCAACGTGAGCCATGAGTTACGTAC
GCCATTGACCGTGTACAGGGTTACCTGGAGATGATGAATGAGCAGCGCC
TGAAGGGCGCGGTACCGGAAAAGCGTTGCAACACATGCGCGAGCAGACC
CAGCGATGGAGAGCTGTGAAGCATTTGCTGACGCTCTGAAATAGA
AGCGCACCGACGATTTGCTCAATGAAAAGTTGATGTCCCGATGATGC
TGGCGCTTGTGAGCGCGAGGCTCAGACTCTGAOTCAGAAAAACAGACA
TTTACCTTTGAGATAGATAACGGCTCAAGGTGTCTGGCAACGAAGATCA
GCTACGCAAGTGGATTTGAACTGGTGTATACGGCTGTGAATCATACAC
CGAAGGCGACGATATCAACGTACGCTGCAACGAGTGGCGACGGTGGC
GAATTTAGCTTTGAAGATAACGGACGGGCTATTGACCGGAGCATATTCC
GGCGCTGACCGAGCGTTTTATCGCTTGTAAAGCGCTTCCCGGCAAA
CCGGCGTACGGGATTAGGTTAGCGATCTGAAACATGCTGTGATCAT
CACGAATGCTGCTGATATTGAGATACATGAGAAAGGACACGCTTT
CAGTTTTGTTATCCCGAAGCTTTAATGCCAAAACAGCGATTAA

Retrieve flanking DNA:
☒ Downstream ☐ Upstream 500 bp [Go](#)

Component: 142 (Escherichia coli K12, complete genome)
Location: 417113..418408
Strand: +
GC content: 53.0%

NCBI Refseq Annotation

NCBI Gene ID: 945044
Name: phoR
Experimental: N
Other notes: synonyms: nmpB, phoR1, R1pho, EG10733, b0400

Domain Architecture:

Pfam (Details)

SMART (Details)

Chromosome View:

Figure 3.12 Protein and gene web page for the MiST database. Basic annotation and sequence data for a protein and its corresponding gene are displayed along with the predicted domain architecture and genome neighborhood.

CHAPTER 4

ONE-COMPONENT REGULATORS DOMINATE SIGNAL TRANSDUCTION IN PROKARYOTES

This chapter is an adapted reproduction of our recent publication in *Trends in Microbiology*, One-component Regulators Dominate Signal Transduction in Prokaryotes (2005) by Ulrich, L.E., Koonin, E.V., and Zhulin, I.B.

Introduction

Two-component systems that link environmental signals to cellular responses are viewed as the primary mode of signal transduction in prokaryotes. By analyzing information encoded in 145 prokaryotic genomes, we found that the majority of signal transduction systems consist of a single protein containing input and output domains but lacking phosphotransfer domains typical of two-component systems. One-component regulators are evolutionary older, more widely distributed among bacteria and archaea, and display a greater diversity of domains than two-component systems.

Signal transduction pathways in prokaryotes regulate cellular functions in response to environmental cues. According to the current view, prokaryotic signal transduction is comprised mostly of two-component regulatory systems which function through phosphotransfer between two key proteins, a sensor histidine kinase and a response regulator (Hoch, 2000; Hoch and Silhavy, 1995; Inouye and Dutta, 2003; Parkinson, 1993; Parkinson and Kofoed, 1992; Stock, et al., 2000). Most experimentally studied histidine kinases are membrane-bound and have an extracellular input domain,

whereas all response regulators are cytosolic. In a typical two-component system, the input domain of the sensor histidine kinase detects the environmental signal (usually a small molecule ligand) thereby activating the histidine kinase domain which autophosphorylates at a specific histidine residue (H) (Figure 4.1a). The phosphoryl group (P) is then transferred to a specific aspartate residue (D) in the receiver domain of a response regulator. Phosphorylation of the response regulator activates the output domain, which initiates the corresponding cellular response. The majority of experimentally characterized two-component systems regulate gene expression at the level of transcription via the DNA-binding helix-turn-helix (HTH) output domains of the response regulators (Hoch, 2000; Stock, et al., 2000). In addition, some response regulators contain enzymatic output domains, such as the di-guanylate cyclase (Paul, et al., 2004).

What is a One-component System?

The modular design of two-component systems is elegant, but does not seem to be the simplest possible solution to the signal transduction challenge – linking environmental stimuli to adaptive responses. A much simpler design of a signal transducer is direct fusion of an input domain to an output domain in a single protein molecule (Figure 4.1A). Indeed, it is well known that transcription of prokaryotic operons is typically regulated by single-molecule repressors and activators which, like most two-component systems, contain a ligand-binding domain and a DNA-binding HTH domain. The LacI lactose operon repressor (Lewis, et al., 1996) and the CAP activator (Kolb, et al., 1993) of *E. coli* are classic examples of such transcriptional regulators. Although these transcriptional regulators are not normally described as signal transduction systems,

we and others noticed that they contain some of the same input and output domains that are typical of two-component signal transduction systems (Anantharaman, et al., 2001; Galperin, et al., 2001; Shu, et al., 2003; Shu and Zhulin, 2002; Taylor and Zhulin, 1999; Zhulin, et al., 2003). For example, PAS (Taylor and Zhulin, 1999) and HTH are input and output domains, respectively, in the two-component system NtrB/NtrC (Weiss, et al., 2002) and in the single-molecule transcriptional regulator RocR (Calogero, et al., 1994) (Figure 4.1B). Using domain database searches, we identified many other combinations of input and output domains as direct fusions in known or predicted regulatory proteins (Figure 4.1B). Thus, one-component systems, which are defined as proteins that contain known or predicted input and output domains but lack histidine kinase and receiver domains, appear to have a repertoire of input and output domains similar to that of two-component systems and therefore might detect similar stimuli and elicit similar cellular responses. Sensory and regulatory properties of some of the one-component systems have been well documented experimentally (Shelver, et al., 1997; Spiro and Guest, 1990; Vannini, et al., 2002).

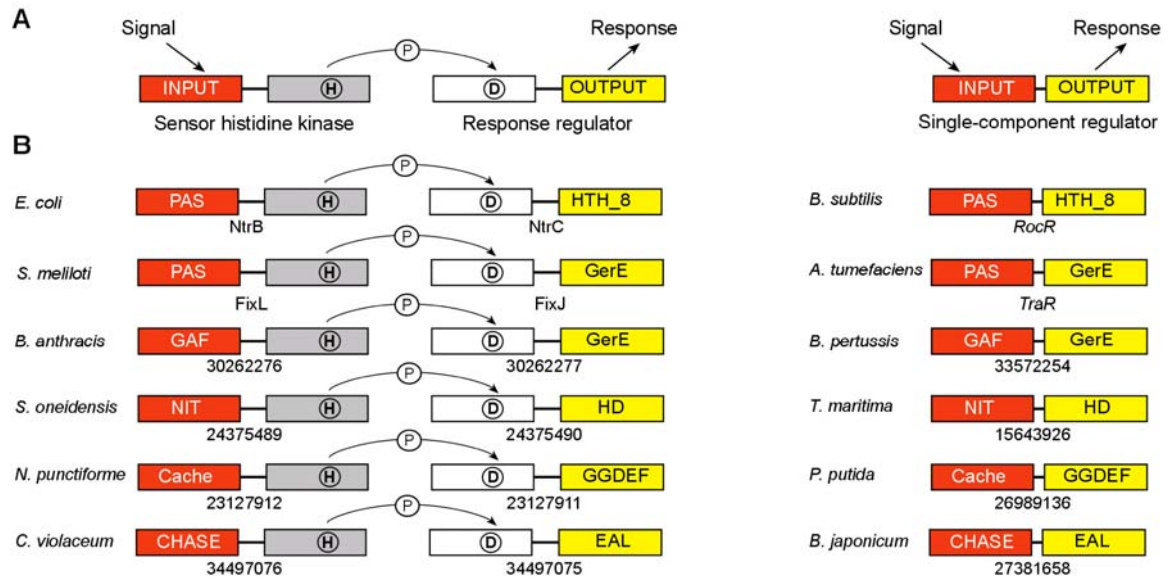


Figure 4.1 Two-component and one-component signal transduction. (A) A prototypical two-component signal transduction system contains input (colored red) and output (colored yellow) domains in two different proteins that communicate via a His-Asp phosphotransfer. A one-component system is a protein that contains input and output domains but lacks His-Asp phosphotransfer domains (colored gray and white). (B) Examples of two-component and one-component systems that utilize the same type of input and output domains. Experimentally studied proteins are identified by name, while proteins predicted from genome sequences are identified by their GenBank ID number.

Detection of Signal Transduction Proteins in Sequenced Genomes

Identification of signal transduction proteins is based on the computational domain analysis of protein sequences. We obtained 145 complete and draft prokaryotic genomes from the National Center for Biotechnology Information (a complete list of genomes is available on our website (<http://genomics.biology.gatech.edu/research/TIM>)). Protein sequences encoded in each genome were searched against the Pfam (Bateman, et al., 2004) and SMART (Letunic, et al., 2004) domain libraries (hidden Markov models)

using the HMMER software package (<http://hmmer.wustl.edu/>) on a parallel Linux cluster. The resulting domain architectures were stored in a MySQL database. Custom Perl scripts were developed to query the database for domains and domain combinations using regular expressions.

Definitions of input and output domains of prokaryotic signal transduction and gene regulation categories are based on curated assignments in the Pfam-A (Bateman, et al., 2004), SMART (Letunic, et al., 2004) and Clusters of Orthologous Groups (COGs) (Tatusov, et al., 2003) resources, and recent genomic surveys (Anantharaman, et al., 2001; Galperin, et al., 2001; Zhulin, et al., 2003). A complete list of input and output domains used in this study can be found on our website (<http://genomics.biology.gatech.edu/research/TIM>). Two-component systems were identified by the presence of the histidine kinase (HATPase_c) and response regulator (response_reg) domains. One-component regulators were identified by the presence of one or more known output domains and the absence of histidine kinase and response regulator domains. Because many input domains involved in prokaryotic signal transduction also participate in other cellular processes (e.g., ligand-binding in transport and metabolism), they were counted only when found in a combination with the histidine kinase (two-component systems) or a known output domain (one-component regulators). In contrast, output domains typically have a single, specific regulatory function (e.g., transcriptional regulation via binding to specific promoters) and therefore all output domains were counted in the domain analysis.

One-component Versus Two-component Systems: a Survey of Bacterial and Archaeal Genomes

Exhaustive database searches and analysis (see previous section) yielded detailed information on the distribution and co-occurrence of input and output domains in 145 complete and draft prokaryotic genomes (see complete results on our website at <http://genomics.biology.gatech.edu/research/TIM>). Strikingly, this analysis detected many more one-component systems (~17,000) than two-component systems (~4,000). Moreover, one-component regulators show much greater diversity of the input and output domain repertoire than two-component systems (Figure 4.2). Many domains are found exclusively in one-component systems, whereas there are no unique input or output domains in two-component systems. The principal type of output activity in both classes of signal transduction systems is regulation of gene expression at the level of transcription: 87% of the known output domains in two-component systems and 84% in one-component regulators are DNA-binding HTH domains. The rest of the output domains are enzymes regulating the level of cyclic nucleotides and protein phosphorylation. The major input activity in both classes is small-molecule-binding: 96% of the known input domains in two-component systems and 93% in one-component regulators are various small-molecule-binding domains. The rest of the input domains are enzymatic and cofactor-containing (mostly redox-responsive) domains and domains involved in protein-protein interactions. Distinct input domains were detected in many but not all of the one-component regulators. However, current computational tools do not detect input domains in protein sequences of several one-component regulators that are

known to carry out specific sensory functions. For example, in the CueR Cu^+ sensor/transcriptional activator, the HTH output domain is fused directly to a simple helix-loop-helix element (undetectable by current computational domain searches), which serves as a metal-sensing (input) domain (Changela, et al., 2003). Therefore, we hypothesize that most if not all one-component regulators detected in our genomic analysis via the identification of an output domain participate in various forms of prokaryotic signal transduction.

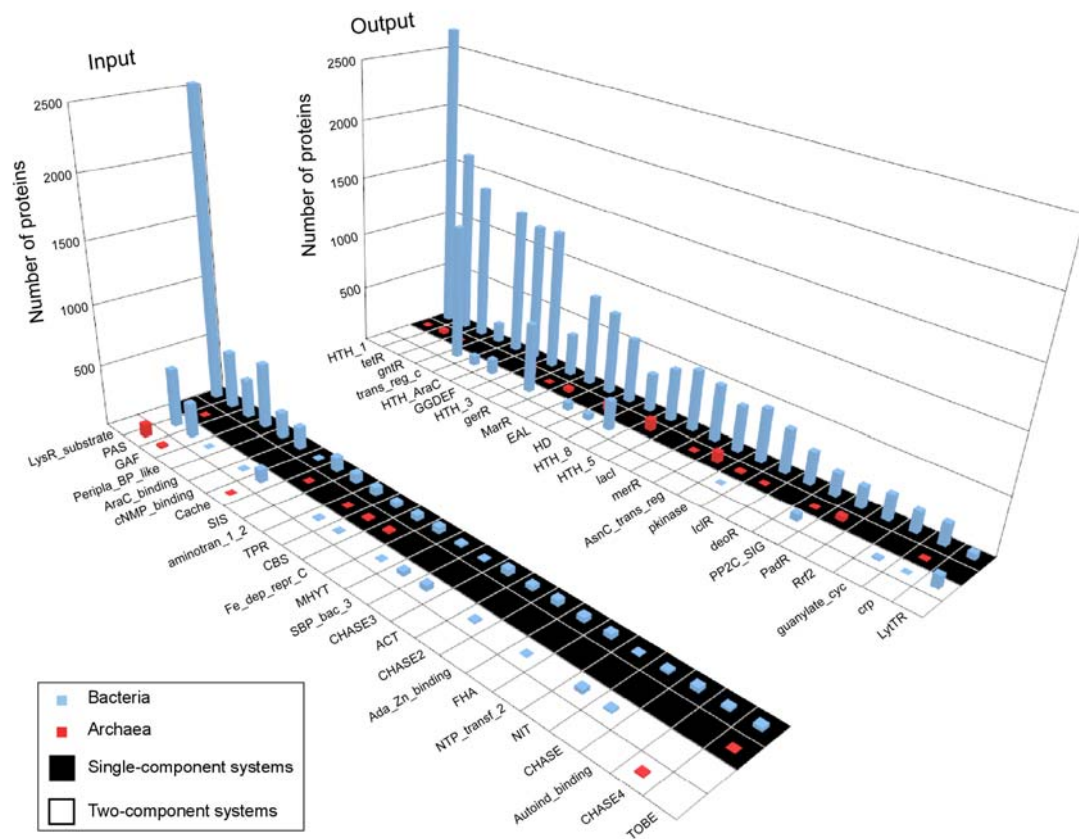


Figure 4.2 Distribution of input and output domains in bacterial and archaeal signal transduction systems. The counts of the twenty-five most abundant input and output domains in bacterial and archaeal one-component and two-component systems are shown. Domain nomenclature is from the curated Pfam-A database (see Appendix A for detailed information).

Genome Size, Lifestyle and Environment Contribute to the Complexity of Signal Transduction

The number of one-component and two-component systems per genome positively correlates with the genome size and, in both cases, is roughly proportional to the square of the total number of genes (Figure 4.3). As shown recently, signal transduction and regulation of gene expression stand out among all functional categories of proteins in showing the steepest dependence on the total number of genes (Konstantinidis and Tiedje, 2004; Van Nimwegen, 2003). This seems to reflect the disproportionate increase in the hierarchical complexity of gene regulation with the increase in genome size, which might ultimately control the maximum achievable genome size, at least in prokaryotes. The results presented here indicate that both one-component and two-component systems contribute to this increase in biological complexity, but given the similar exponents of the plots (Figure 4.3), the contribution of the more abundant one-component regulators is greater. Significant deviations from this general trend seem to reflect particular biological phenomena as well as environmental conditions in a microbial habitat. For example, the genomes of the marine cyanobacterium *Trichodesmium erythraeum* and the soil α -proteobacterium *Sinorhizobium meliloti* are comparable in size: 7.7 and 6.7 Mb, respectively. However, there are 69 one-component regulators encoded in the former (unusually few) versus 390 in the latter (unusually many). The difference in the number of two-component systems in the two genomes is much less noticeable and, in fact, not significant: 35 and 40, respectively. Both bacterial species have a versatile metabolism, which is reflected by

their large genome size. However, *S. meliloti* has a complex developmental program (Galibert, et al., 2001) and experiences significant fluctuation of various physico-chemical parameters in its microenvironments (soil, rhizosphere, plant root interior). In contrast, *T. erythraeum* does not undergo developmental changes typical of other nitrogen-fixing cyanobacteria (heterocyst formation) and lives under more or less constant environmental conditions (upper levels of tropical oceans) (Staal, et al., 2003).

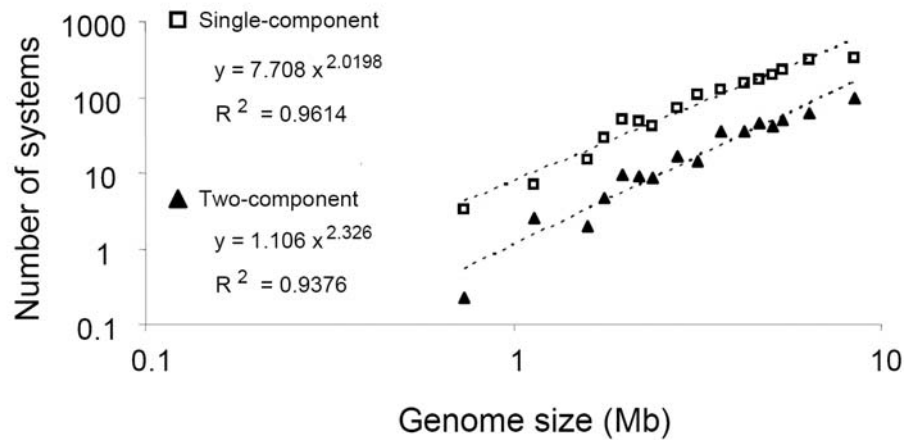


Figure 4.3 Dependence of the number of one-component and two-component signal transduction systems on the genome size. The plot is in a double logarithmic scale. One hundred forty-five genomes were ranked by size and split into 16 size classes. Each point indicates the average number of genes for one-component or two-component signal-transduction systems in the respective class.

One-component Systems as the Primordial Form of Prokaryotic Signal Transduction

Three lines of evidence suggest that one-component regulators are evolutionary precursors of the two-component systems. Firstly, the modular design of one-component systems is obviously simpler than that of two-component systems. Secondly, as shown above, the domain repertoire of one-component regulators is considerably more diverse than that of two-component systems. Finally, one-component regulators are more widely distributed among prokaryotes than two-component systems: with the exception of some parasites with highly degraded genomes, such as mycoplasmas, all prokaryotes encode a substantial diversity of one-component regulators. By contrast, two-component systems are (nearly) missing in many species, particularly, among archaea (Figure 4.2). Furthermore, archaeal two-component systems do not seem to form a coherent lineage (data not shown) and probably have been acquired from bacteria via horizontal gene transfer as previously suggested (Koretke, et al., 2000). Therefore it appears most likely that the last common ancestor of archaea and bacteria (i.e., the last common ancestor of all modern life forms) did not have two-component systems, but probably encoded several one-component regulators. Two-component systems appear to be a subsequent bacterial innovation, which emerged through insertion of histidine kinase domains and receiver domains into one-component regulators. If one-component regulatory systems comprise such a straightforward solution to the requirements of prokaryotes for signal transduction, then, what is the advantage of two-component systems? We believe that this has to do with the fact that one-component regulators detect stimuli (including environmental cues, such as gases, light, and various small molecules) almost exclusively

in the cytosol. We scanned all 25,303 protein sequences identified as components of signal transduction systems in 145 prokaryotic genomes for the presence of transmembrane regions using the DAS method (Cserzo, et al., 2004) and found that 97% of one-component regulators that contain an HTH domain do not have transmembrane regions and therefore are predicted to be cytosolic proteins. In contrast, more than 73% of the sensor histidine kinases were predicted to be membrane-associated based on the presence of one or more transmembrane regions. Thus, the fundamental difference in the sensing mode between the one-component and two-component systems is intracellular versus extracellular detection of stimuli, respectively. Extracellular sensing provides a microbe with an obvious advantage compared to exclusive intracellular sensing. However, because more than 80% of signal transduction pathways involve DNA-binding, arrangement of single-molecule regulators in the membrane would place major constraints on their ability to interact with their targets in genomic DNA. A straightforward and efficient solution to this problem is dividing the signal transduction system into two proteins, a membrane-bound sensor and a soluble, cytosolic DNA-binding regulator, which are linked via a phosphotransfer relay. Hence the emergence of the two-component signals transduction systems.

Conclusions

The availability of a large number of sequenced prokaryotic genomes allowed us to reveal the dominance of one-component signal-transduction systems and the apparent ancestor-descendant relationship between them and the two-component systems in prokaryotes. It has been noticed previously that some of the transcriptional regulators in prokaryotes possess sensory properties. However, to our knowledge, the fact that two-

component signal transduction systems utilize a subset of the input and output domains that are present in one-component regulators so far has not been recognized, and neither was the extraordinary combinatorial diversity of one-component regulators.

Acknowledgements

The literature on signal transduction in prokaryotes is vast. We extend our apologies and appreciation to all colleagues whose work is not cited here solely due to space limitations. We thank Eugene Koonin for his participation in co-authoring this manuscript. We thank Yuri Wolf for assistance with the genome size dependence analysis and Michael Galperin, Susan Golden and Sydney Kustu for helpful discussions. This work was supported by research grants GM72285 from National Institutes of Health and EIA-0219079 from National Science Foundation (to I.B.Z.). L.E.U. was supported by IGERT-0221600 grant from National Science Foundation.

CHAPTER 5

FOUR-HELIX BUNDLE: A UBIQUITOUS SENSORY MODULE IN PROKARYOTIC SIGNAL TRANSDUCTION

This chapter is an adapted reproduction of our recent publication in *Bioinformatics*, Four-Helix Bundle: a Ubiquitous Sensory module in Prokaryotic Signal Transduction (2005) by Ulrich, L.E. and Zhulin, I.B.

Abstract

Transmembrane chemoreceptors in *Escherichia coli* utilize ligand-binding domains for detecting various external signals. The structure of this domain in the *E. coli* aspartate receptor, Tar, is known and its signal transduction mechanism is under investigation. Current domain models for this important sensory module are inaccurate and therefore cannot reveal the distribution of this domain within the current genomic landscape. We carried out sensitive and exhaustive PSI-BLAST searches initiated with the sequence corresponding to a known structure of the four-helix, ligand-binding domain of the aspartate chemoreceptor. From the resulting sequences, we built a multiple sequence alignment for this domain family, which confirmed that the current TarH model is erroneous and fails to detect most of the domain homologs. In the process, we developed a technique that visualizes the secondary structure prediction of each protein sequence in order to improve the multiple sequence alignment. We found that the four-helix up-and-down bundle represents a large domain family and includes representatives

of all major classes of prokaryotic signal transduction, namely histidine kinases, diguanylate cyclases, and chemotaxis receptors.

Introduction

E. coli detects several attractant and repellent compounds using transmembrane chemoreceptors (also known as methyl-accepting chemotaxis proteins or MCPs) that transmit information to flagellar motors via a signal transduction pathway (Falke and Hazelbauer, 2001; Sourjik, 2004). Many other motile bacteria and archaea have homologous pathways for regulating motility, and MCPs comprise a large protein superfamily (Zhulin, 2001). All MCPs have a conserved signaling domain and variable sensory domains. The structure of the sensory (ligand-binding) domain of the *E. coli* chemoreceptor, Tar, has been solved (Bowie, et al., 1995; Milburn, et al., 1991) and revealed an antiparallel four-helix bundle. Three other transmembrane chemoreceptors of *E. coli* – Tsr, Trg, and Tap – were proposed to have homologous ligand-binding domains based on limited sequence similarity (Zhulin, 2001). More recently, a good quality 3D model of the Trg ligand-binding domain has been built using the Tar crystal structure (Peach, et al., 2002). The model was consistent with previously obtained mutational data and is strongly supported by selective formation of disulfides between introduced cysteines (W. Lai and G.L. Hazelbauer, personal communication).

Both leading domain databases, Pfam (Bateman, et al., 2004) and SMART (Letunic, et al., 2004), contain a domain model called TarH (accession number SM00319), which has been built from the multiple alignment of protein sequences corresponding to the ligand-binding domains of Tar, Tsr, Trg, and Tap from *E. coli* and closely related enteric bacteria, and three sequences of predicted ligand-binding domains

from the *Bacillus subtilis* chemoreceptors McpA, McpB, and TlpB. The Pfam and SMART TarH models recognize 79 and 42 sequences in the non-redundant database, respectively. Because McpA, McpB, and TlpB contain a Cache domain (Anantharaman and Aravind, 2000) as their core ligand-binding element (which is not present in *E. coli* chemoreceptors), we questioned the appropriateness of their alignment to the *E. coli* ligand-binding domains in the TarH model. Furthermore, the TarH alignment contains several large gaps introduced by including the *B. subtilis* sequences and is of an overall poor quality.

Methods

We began the domain analysis by initiating an exhaustive and sensitive PSI-BLAST search (Altschul, et al., 1997) against the non-redundant database (May 13, 2005) with residues 34-190 of the ligand-binding domain from the *E. coli* Tar chemoreceptor. This region is flanked by two transmembrane helices and encompasses the entire solved structure (residues 39-178) of this sensory domain. PSI-BLAST searches (E-value cutoff 0.01) were continued until no new homologs were identified. Because this domain is highly alpha-helical, a search was stopped upon hits to unrelated random coils such as those found in myosin- or laminin-like proteins. Partial and duplicate sequences were excluded from this analysis.

Multiple sequence alignments were constructed using the MUSCLE (Edgar, 2004) and ClustalW (Thompson, et al., 1994) programs and manually adjusted. While editing this alignment we developed VISSA (Visualization of Secondary Structure elements for Improving multiple Alignments), a technique that involves visualization of predicted secondary structure elements (using color) on each sequence within an MSA.

This visualization allows a rapid check for consistency between the MSA features (e.g. gaps) and the predicted secondary structure. The VISSA technique consists of two steps: *data processing* and *visualization* (Figure 5.1):

- 1) *Data processing*: A ClustalW-formatted MSA is loaded into memory and the secondary structure of each sequence is predicted using PSIPRED (Jones, 1999). The alignment, prediction results, and various metadata are stored in an XML file. In addition to the alignment, this XML file contains the predicted secondary structure and confidence values for each sequence in the MSA.
- 2) *Visualization*: The results within the XML data file are then visualized by creating an HTML document, which has the background of alpha helices and beta strands shaded red and blue, respectively, in proportion to the confidence of each prediction – darker shades representing a higher confidence value. The text color of loop regions is shaded in a similar fashion.

Custom Perl scripts handle the data processing and visualization steps. The technical implementation was meant to be as abstract as possible. An XML document specification is provided which describes how to organize the data for a multiple alignment in conjunction with its secondary structure. Given this abstraction, the individual may then predict the secondary structure using any particular tool and visualize this data with VISSA provided that the data conforms to this XML specification. Similarly, others may wish to visualize this information differently or process this data with another program. In such cases, the intermediate XML document is a structured, machine and human readable data file enabling the user to easily process the data. Sequence conservation

information and the visualized secondary structure of the four-helix bundle domains enabled us to produce a much better quality alignment.

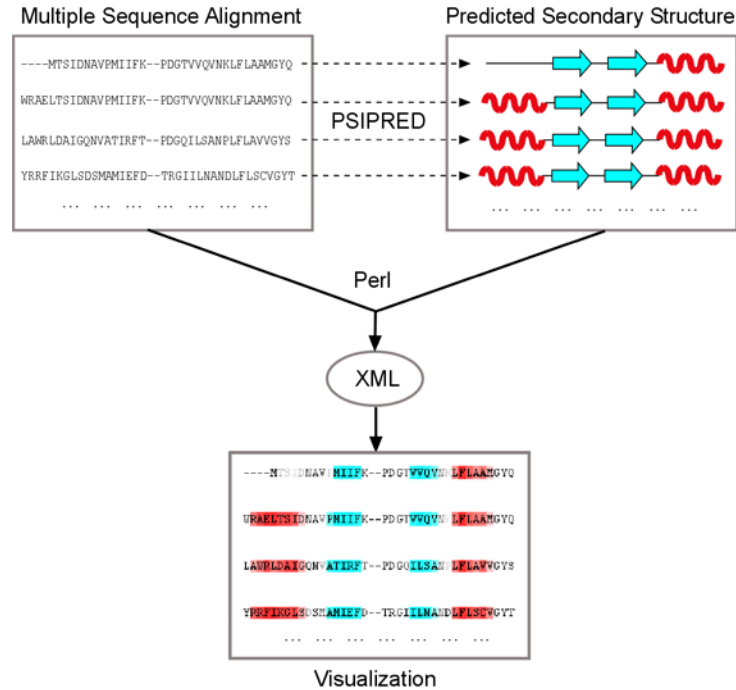


Figure 5.1 Overview of the VISSA process. For each sequence in a multiple alignment, its secondary structure is predicted and both the alignment and structural information stored in an XML document. This XML data is then visualized by coloring/shading the amino acids that correspond to the predicted secondary structure elements.

Domain architectures were predicted using the HMMER software package (Eddy, 1998) and the Pfam (version 17.0) and SMART (version 4.0) domain libraries. Signal peptides and transmembrane regions and their topology were assigned using Phobius

(Kall, et al., 2004). The secondary structure of each homolog was predicted using PSIPRED (Jones, 1999). Sequence logos based on the information content were created with the Berkeley weblogo server (Crooks, et al., 2004). The three-dimensional structure of the Tar ligand-binding domain (PDB accession 2ASR) was visualized using the PyMOL Molecular Graphic System (www.pymol.org).

Results and Discussion

Exhaustive PSI-BLAST searches resulted in the detection of more than 700 copies of the domain, which we termed 4HB_MCP for a four-helix bundle found predominantly in the N-terminal region of MCPs. These results are in striking contrast with the performance of the current TarH model in domain databases, which recognizes relatively few sequences. None of the searches detected the 4HB_MCP domain in the McpA, McpB, and TlpB proteins that were used in the seed alignment for the TarH model, thus confirming its deficiency.

A

Ecol_2506837_36-190
Ecar_50120625_36-192
Bjap_27376721_106-258
Sent_62180191_48-202
Dhaf_23120317_33-189
Ecar_50122563_29-187
Reut_46131863_31-193
Cvio_34102638_20-177
Reut_53761244_29-187
Ecar_50119143_33-189
Rleg_4973017_2-195
Gsul_39996396_31-188
Xcam_21231323_1-138
Bpse_53721497_32-189
Bfun_48788500_33-185
Bfun_48787377_20-177
Psysr_46188178_11-168
Rshp_7532754_8-185
Bfun_48782649_13-171
Naro_48849030_29-187
Neur_30249816_32-185
Dvul_46578600_31-190
Pflu_48732089_27-183
Psysr_23472827_32-190
Iloi_56459719_31-188
Bsub_16080422_28-184
Exsp_46114138_32-187
Bcer_52142147_38-193
Wsuc_34557328_34-190
Vvul_27358478_31-189

consensus/70%

SQKSPVVSNOI~~REQO~~GELTSTWDLMLQTRINLSRAVRMMDDSS-14--QOSNAKVELLDSARKTLAQAA~~THKK~~FKSM-6-PLEPMVATSRNIDEKYKNYYALTE~~LID~~LDYGN-10----PGAYFAQPTQGMONAMGEAFAQYALSSEKLYRDI~~VT~~DNADDYRFA-
DKDIFSSQTQVINOGRSELD~~SA~~NSYLLQTRNTLN~~RAGT~~RFALDVS-14--VGGEKELLTSAEKOLAVANDYFIRYEKMP-Q-6-ARQDSDISRGVKNYVAINAALTE~~LIC~~QFINA-GE-10----FKKFI~~EQ~~PTQRFQDNFEKAYVYVKAENDKLYQAGIAKNDAA~~YDS~~A-
--QCVRLQTN~~YR~~HMQIDSASAAVNI~~GR~~VNALIYAIVMES~~SGI~~YIM-14--PAKVQFAD~~EL~~VKCSGELTAVMMRWGQTV--6---HDDLEQFEAFQORVMQIDFRLE~~L~~VRR-GL-10-ISPPAAAREWDDNQ-ANRTVRSKLNADLEALQ~~RSY~~DKRAREADQ~~LA~~DENR-
--LQADRDQ~~RD~~VTEIQVRMG~~LS~~NSANHLRTARINMIHAGAAS-14--MDEMKANIAA~~ETRI~~KQSDGDFNAYMSR-VK-6--TPADDALDNE~~L~~NARYATYINGLQPM~~L~~FKFKN-GM-10-FEATINHENEQAKOLDAAYNHVL~~L~~KAL~~EL~~TERARLLSEQA~~YQ~~RT--
EIEQHRARGALQSLQ~~DSL~~RGLRLIKTVSDAYGLDVDDTFFRV~~R~~-14--VQDEGVDRVDRARARIDAAWRELDALP-H-6-SPREQQQLNAAQAARRTADAAQELRA~~L~~ILRL-RD-10-LLALGRFAD~~TR~~LYPAIDF~~IL~~TRMQQ~~LS~~DELEIQADAVVRADIVRSERV-
EYQLNQVSSSQOQM~~QE~~FLKERLASD~~R~~HATLVAGVQ~~RS~~MAVAR-14--DSLVELFAAENTRASKE~~SG~~KRQEDFASLI-S-6-TPEEKALFDKVGEYRQSYIKKRDAIITEKGA-GN-10-FDRARTLFDNEFVPASNGYLASVEALRD~~HQ~~QRASIDOMGKNI~~IN~~AGASRS~~GD~~
VKSINDANDR~~AS~~FA~~SV~~QGFNAQ~~LL~~SYEVRTA~~I~~ERRAILARD~~V~~N-14--QDRATIKAEVTTVHDDQ~~ARI~~AKLT~~K~~MAAD-6-SREARTLVAKIAEVETKYSV~~AL~~GIVELASQ-GK-10-REEAVAKMNADCRPLLAAL~~I~~HAHDYR~~DF~~TSKHS~~ED~~LVTQAAADY~~AM~~QR-
YNKLQTIQNNITEIKEDRY~~KE~~VL~~SQ~~RIALN~~L~~LYISRGVRD~~GV~~L-14--QKVEQQIRNVEVL~~R~~ANNRADL~~DK~~MEPM-S-6-TPEGRALFAKIRAAQDSQRPLFE~~PL~~YGLMRS-HQ-10-TDAARDML~~EN~~QFAP~~T~~NNAFISALLSLRD~~R~~Q~~QS~~RLDK~~Q~~AE~~MD~~SSQA-
YNRLSAIERETN~~L~~MLKDALPGLN~~ST~~GTIRGAWGEVYVLAWETVK-14--ESORQGYQQQAEV~~RQ~~RLDRLEQO~~VE~~ST-T-6-RDNBRATFN~~OV~~YKAARS~~RY~~DQ~~LA~~LPLTQ~~SA~~MSK-G-10-GEAAEALRG~~EA~~NOHWA~~EV~~RR~~LA~~Q~~TV~~LDVDDNNAVNEKA~~AG~~HNIA~~AV~~SSAK-
--KLSGEQDSARSIVKDVFPQVDANNLIDNV~~TS~~LIVAYQRLML-14-SVQIQTNVTRVNE~~RQ~~EIRLLDKLERQTV-6-EERSVTQLRAIRAIT~~EE~~LKSGDKLISEVVA-GN-10-REAAIEEFNNNLNV~~VQ~~RQ~~Y~~RD~~AV~~Q~~KL~~VNNQDDAD~~MT~~SV~~EA~~MAEVYSNTR-
TEEMALINDKL~~GA~~MNDVNSVKORFAIN~~YR~~SGSVH~~DR~~AIR~~AD~~RVTL-14--DDERKTAEALIGKLAAS~~YA~~ENEK~~R~~MMAD~~MA~~VS-6-TQE~~KT~~LILSEADIQAKANPLVAQ~~LI~~ALQ~~EQ~~-GD-10-GEAARKILLEQARPAFV~~AW~~LGA~~IN~~K~~FI~~DYQ~~AL~~NK~~SI~~IGGEV~~RS~~SA~~SG~~FK-
--NRMATINTD~~LM~~VVKDRM~~KE~~AEITFGISSQIN~~V~~ARALRN~~LT~~L-14--AEVQKEIARINEASV~~SK~~SMDS~~EL~~SKSI-T-6-SEEGAKKLKAVEASRAA~~YR~~EDLLK~~L~~VEY~~TR~~A-GN-10-KSAQKMLFGSGYRERQ~~RSY~~FD~~AV~~GLTQYQAKILAVSGKEAEQT~~TV~~SSR-
-----MHLQDMDTDSVTAI~~Q~~LNRLN~~VL~~-14-QEDNLRF~~AA~~LIDKRAKAYEQ~~TR~~Q~~TY~~LYFS-A-6-SPEAQARRDKDAASAAAKANAAQVAELGLA-SK-10-SDEALAMLMQQAAPATEAWQSALAEYSALQ~~RR~~K~~RA~~K~~TA~~YEDATAAMARG-
IYER~~RR~~YTAASYSYTVNTVSE~~TV~~VLDDAQRA~~FD~~SM~~LL~~LNQO~~VF~~-14-ADQAKALEPRIAQA~~RR~~VEDAQ~~FA~~KAYETLL-S-6-NDKKKALLAADRARVSQ~~LD~~AVREN~~V~~IALR-GR-10-QKEAGELMGT~~RT~~MTLEAQQNAALAAH~~AE~~NV~~DL~~GQAGSNEAK~~DI~~TDRA-
--GMSRNNHALSDTFNAPSAVDIGNAEL~~Y~~AERERIALDRAAF-14-TPEEA~~PT~~LERAR~~GR~~MATDMW~~KQ~~MDLR-R-6-EPEDERLAQDVVSRREALHQ~~LD~~DAFAALFAA-ND-10-QAKLV~~DG~~A-KRLQVAYNDLANADALR~~KY~~QTS~~AK~~EGYDAE~~SS~~FE--
QLGMRKANDELAYAYS~~NQ~~LIAAIAAGEANLS~~TV~~ARLSLD~~R~~ALL-14-SPVD~~PL~~LIARTQ~~RL~~DRAD~~R~~YALP-H-6-ESEERVLADRVNAA~~RT~~ALLQVQ~~FS~~IEALAKGE-10-HERADAI~~VM~~K~~MT~~MSLSLAL~~T~~NSADAL~~TD~~WQKAHQRQAFAD~~Q~~RL~~HD~~R--
YVQIGHL~~RV~~AEQNI~~EN~~SLSPVQVDDIQIALH~~AR~~LESIRMLA-14-PSVHASAEAKVREATEALRANS~~DF~~QKILLS-6-GEACAPQFEAN~~KN~~MGVITDGKLVQ~~SV~~ALDS--AD-10-HERAVSLANGEQALKAAYQKAL~~AD~~TRGRGHNAEAV~~VS~~GRDA~~YV~~DH-
LRDLDQIRASLDDIVHTKVKQVEMTYQ~~LI~~ENRLK~~Q~~REIRN~~LT~~L-14-KEERRAID~~DR~~LATASAG~~EA~~FAALE-A-6-DAETRARI~~AE~~VQAEKERIART~~DE~~KAIEMAR~~ML~~-GL-10-GYEGFTIV~~VT~~QGEQW~~LA~~METRLSALLAH~~HT~~Q~~OL~~TDASAEAQ~~RG~~Q~~IS~~R-
Bfun_48788500_33-185
Bfun_48787377_20-177
Psysr_46188178_11-168
Rshp_7532754_8-185
Bfun_48782649_13-171
Naro_48849030_29-187
Neur_30249816_32-185
Dvul_46578600_31-190
Pflu_48732089_27-183
Psysr_23472827_32-190
Iloi_56459719_31-188
Bsub_16080422_28-184
Exsp_46114138_32-187
Bcer_52142147_38-193
Wsuc_34557328_34-190
Vvul_27358478_31-189

..phtphtphtpph.pp.hsshthhsphpthtthttht.hphhh-14-.tthpph.tphpphtphtphtphtpphphh.t-6-s.ppcshhsphpthtthhphhtphtphttt.Gp-10-hptAhshh.ppt.h.phttshphtthtthpphtphtphtppshpshps.

B

Ecol_2506837_36-190
Ecar_50120625_36-192
Bjap_27376721_106-258
Sent_62180191_48-202
Dhaf_23120317_33-189
Ecar_50122563_29-187
Reut_46131863_31-193
Cvio_34102638_20-177
Reut_53761244_29-187
Ecar_50119143_33-189
Rleg_4973017_2-195
Gsul_39996396_31-188
Xcam_21231323_1-138
Bpse_53721497_32-189
Bfun_48788500_33-185
Bfun_48787377_20-177
Psysr_46188178_11-168
Rshp_7532754_8-185
Bfun_48782649_13-171
Naro_48849030_29-187
Neur_30249816_32-185
Dvul_46578600_31-190
Pflu_48732089_27-183
Psysr_23472827_32-190
Iloi_56459719_31-188
Bsub_16080422_28-184
Exsp_46114138_32-187
Bcer_52142147_38-193
Wsuc_34557328_34-190
Vvul_27358478_31-189

SQKSPVVSNOI~~REQO~~GELTSTWDLMLQTRINLSRAVRMMDDSS-14--QOSNAKVELLDSARKTLAQAA~~THKK~~FKSM-6-PLEPMVATSRNIDEKYKNYYALTE~~LID~~LDYGN-10----PGAYFAQPTQGMONAMGEAFAQYALSSEKLYRDI~~VT~~DNADDYRFA-
DKDIFSSQTQVINOGRSELD~~SA~~NSYLLQTRNTLN~~RAGT~~RFALDVS-14--VGGEKELLTSAEKOLAVANDYFIRYEKMP-Q-6-ARQDSDISRGVKNYVAINAALTE~~LIC~~QFINA-GE-10----FKKFI~~EQ~~PTQRFQDNFEKAYVYVKAENDKLYQAGIAKNDAA~~YDS~~A-
--QCVRLQTN~~YR~~HMQIDSASAAVNI~~GR~~VNALIYAIVMES~~SGI~~YIM-14--PAKVQFAD~~EL~~VKCSGELTAVMMRWGQTV--6---HDDLEQFEAFQORVMQIDFRLE~~L~~VRR-GL-10-ISPPAAAREWDDNQ-ANRTVRSKLNADLEALQ~~RSY~~DKRAREADQ~~LA~~DENR-
--LQADRDQ~~RD~~VTEIQVRMG~~LS~~NSANHLRTARINMIHAGAAS-14--MDEMKANIAA~~ETRI~~KQSDGDFNAYMSR-VK-6--TPADDALDNE~~L~~NARYATYINGLQPM~~L~~FKFKN-GM-10-FEATINHENEQAKOLDAAYNHVL~~L~~KAL~~EL~~TERARLLSEQA~~YQ~~RT--
EIEQHRARGALQSLQ~~DSL~~RGLRLIKTVSDAYGLDVDDTFFRV~~R~~-14--VQDEGVDRVDRARARIDAAWRELDALP-H-6-SPREQQQLNAAQAARRTADAAQELRA~~L~~ILRL-RD-10-LLALGRFAD~~TR~~LYPAIDF~~IL~~TRMQQ~~LS~~DELEIQADAVVRADIVRSERV-
EYQLNQVSSSQOQM~~QE~~FLKERLASD~~R~~HATLVAGVQ~~RS~~MAVAR-14--DSLVELFAAENTRASKE~~SG~~KRQEDFASLI-S-6-TPEEKALFDKVGEYRQSYIKKRDAIITEKGA-GN-10-FDRARTLFDNEFVPASNGYLASVEALRD~~HQ~~QRASIDOMGKNI~~IN~~AGASRS~~GD~~
VKSINDANDR~~AS~~FA~~SV~~QGFNAQ~~LL~~SYEVRTA~~I~~ERRAILARD~~V~~N-14--QDRATIKAEVTTVHDDQ~~ARI~~AKLT~~K~~MAAD-6-SREARTLVAKIAEVETKYSV~~AL~~GIVELASQ-GK-10-REEAVAKMNADCRPLLAAL~~I~~HAHDYR~~DF~~TSKHS~~ED~~LVTQAAADY~~AM~~QR-
YNKLQTIQNNITEIKEDRY~~KE~~VL~~SQ~~RIALN~~L~~LYISRGVRD~~GV~~L-14--QKVEQQIRNVEVL~~R~~ANNRADL~~DK~~MEPM-S-6-TPEGRALFAKIRAAQDSQRPLFE~~PL~~YGLMRS-HQ-10-TDAARDML~~EN~~QFAP~~T~~NNAFISALLSLRD~~R~~Q~~QS~~RLDK~~Q~~AE~~MD~~SSQA-
YNRLSAIERETN~~L~~MLKDALPGLN~~ST~~GTIRGAWGEVYVLAWETVK-14--ESORQGYQQQAEV~~RQ~~RLDRLEQO~~VE~~ST-T-6-RDNBRATFN~~OV~~YKAARS~~RYDQ~~LA~~LPLTQ~~SA~~MSK-G-10-GEAAEALRG~~EA~~NOHWA~~EV~~RR~~LA~~Q~~TV~~LDVDDNNAVNEKA~~AG~~HNIA~~AV~~SSAK-
--KLSGEQDSARSIVKDVFPQVDANNLIDNV~~TS~~LIVAYQRLML-14-SVQIQTNVTRVNE~~RQ~~EIRLLDKLERQTV-6-EERSVTQLRAIRAIT~~EE~~LKSGDKLISEVVA-GN-10-REAAIEEFNNNLNV~~VQ~~RQ~~Y~~RD~~AV~~Q~~KL~~VNNQDDAD~~MT~~SV~~EA~~MAEVYSNTR-
TEEMALINDKL~~GA~~MNDVNSVKORFAIN~~YR~~SGSVH~~DR~~AIR~~AD~~RVTL-14--DDERKTAEALIGKLAAS~~YA~~ENEK~~R~~MMAD~~MA~~VS-6-TQE~~KT~~LILSEADIQAKANPLVAQ~~LI~~ALQ~~EQ~~-GD-10-GEAARKILLEQARPAFV~~AW~~LGA~~IN~~K~~FI~~DYQ~~AL~~NK~~SI~~IGGEV~~RS~~SA~~SG~~FK-
--NRMATINTD~~LM~~VVKDRM~~KE~~AEITFGISSQIN~~V~~ARALRN~~LT~~L-14--AEVQKEIARINEASV~~SK~~SMDS~~EL~~SKSI-T-6-SEEGAKKLKAVEASRAA~~YR~~EDLLK~~L~~VEY~~TR~~A-GN-10-KSAQKMLFGSGYRERQ~~RSY~~FD~~AV~~GLTQYQAKILAVSGKEAEQT~~TV~~SSR-
-----MHLQDMDTDSVTAI~~Q~~LNRLN~~VL~~-14-QEDNLRF~~AA~~LIDKRAKAYEQ~~TR~~Q~~TY~~LYFS-A-6-SPEAQARRDKDAASAAAKANAAQVAELGLA-SK-10-SDEALAMLMQQAAPATEAWQSALAEYSALQ~~RR~~K~~RA~~K~~TA~~YEDATAAMARG-
IYER~~RR~~YTAASYSYTVNTVSE~~TV~~VLDDAQRA~~FD~~SM~~LL~~LNQO~~VF~~-14-ADQAKALEPRIAQA~~RR~~VEDAQ~~FA~~KAYETLL-S-6-NDKKKALLAADRARVSQ~~LD~~AVREN~~V~~IALR-GR-10-QKEAGELMGT~~RT~~MTLEAQQNAALAAH~~AE~~NV~~DL~~GQAGSNEAK~~DI~~TDRA-
--GMSRNNHALSDTFNAPSAVDIGNAEL~~Y~~AERERIALDRAAF-14-TPEEA~~PT~~LERAR~~GR~~MATDMW~~KQ~~MDLR-R-6-EPEDERLAQDVVSRREALHQ~~LD~~DAFAALFAA-ND-10-QAKLV~~DG~~A-KRLQVAYNDLANADALR~~KY~~QTS~~AK~~EGYDAE~~SS~~FE--
QLGMRKANDELAYAYS~~NQ~~LIAAIAAGEANLS~~TV~~ARLSLD~~R~~ALL-14-SPVD~~PL~~LIARTQ~~RL~~DRAD~~R~~YALP-H-6-ESEERVLADRVNAA~~RT~~ALLQVQ~~FS~~IEALAKGE-10-HERADAI~~VM~~K~~MT~~MSLSLAL~~T~~NSADAL~~TD~~WQKAHQRQAFAD~~Q~~RL~~HD~~R--
YVQIGHL~~RV~~AEQNI~~EN~~SLSPVQVDDIQIALH~~AR~~LESIRMLA-14-PSVHASAEAKVREATEALRANS~~DF~~QKILLS-6-GEACAPQFEAN~~KN~~MGVITDGKLVQ~~SV~~ALDS--AD-10-HERAVSLANGEQALKAAYQKAL~~AD~~TRGRGHNAEAV~~VS~~GRDA~~YV~~DH-
LRDLDQIRASLDDIVHTKVKQVEMTYQ~~LI~~ENRLK~~Q~~REIRN~~LT~~L-14-KEERRAID~~DR~~LATASAG~~EA~~FAALE-A-6-DAETRARI~~AE~~VQAEKERIART~~DE~~KAIEMAR~~ML~~-GL-10-GYEGFTIV~~VT~~QGEQW~~LA~~METRLSALLAH~~HT~~Q~~OL~~TDASAEAQ~~RG~~Q~~IS~~R-
Bfun_48788500_33-185
Bfun_48787377_20-177
Psysr_46188178_11-168
Rshp_7532754_8-185
Bfun_48782649_13-171
Naro_48849030_29-187
Neur_30249816_32-185
Dvul_46578600_31-190
Pflu_48732089_27-183
Psysr_23472827_32-190
Iloi_56459719_31-188
Bsub_16080422_28-184
Exsp_46114138_32-187
Bcer_52142147_38-193
Wsuc_34557328_34-190
Vvul_27358478_31-189~~

Figure 5.2 Alignment of representative members of the 4HB_MCP domain. **(A)** An alignment of thirty members of the 4HB_MCP domain from the seed alignment. Conserved residues and their positions are colored with the ClustalX scheme using Jalview (Clamp, et al., 2004): orange – glycine (G), yellow – proline (P), blue – small and hydrophobic amino acids (A, V, L, I, M, F, W), green – hydroxyl and amine amino acids (S, T, N, Q), red – charged amino acids (D, E, R, K), cyan – histidine (H) and tyrosine (Y). **(B)** An alignment of the same thirty members illustrating the VISSA visualization. Regions containing predicted alpha helices and beta sheets have the background shaded red and blue, respectively. The intensity of each shade is directly proportional to the confidence of a given prediction – a darker intensity representing a higher confidence. The identifier for each sequence consists of a species abbreviation, the GenBank identifier, and the coordinates of the sequence. Species abbreviations are as follows: Bcer, *Bacillus cereus*; Bfun, *Burkholderia fungorum*; Bjap, *Bradyrhizobium japonicum*; Bpse, *Burkholderia pseudomallei*; Bsub, *Bacillus subtilis*; Cvio, *Chromobacterium violaceum*; Dhaf, *Desulfitobacterium hafniense*; Dvul, *Desulfovibrio vulgaris*; Ecar, *Erwinia carotovora*; Ecol, *Escherichia coli*; Exsp, *Exiguobacterium sp.*; Gsul, *Geobacter sulfurreducens*; Iloi, *Idiomarina loihiensis*; Naro, *Novosphingobium aromaticivorans*; Neur, *Nitrosomonas europaea*; Pflu, *Pseudomonas fluorescens*; Psyr, *Pseudomonas syringae*; Reut, *Ralstonia eutropha*; Rleg, *Rhizobium leguminosarum*; Rsph, *Rhodobacter sphaeroides*; Sent, *Salmonella enterica*; Vvul, *Vibrio vulnificus*; Xcam, *Xanthomonas campestris*; Wsuc, *Wolinella succinogenes*.

Representative and complete multiple sequence alignments of the 4HB_MCP domain were constructed and edited using the VISSA technique, which revealed several inconsistent features in the alignment (Appendix B, Figure B.1). For example, gaps have been placed inside predicted alpha helices and structural elements have been misaligned in several sequences. Efficient visualization of anomalies allowed us to quickly edit the alignment for consistency in both sequence and secondary structure conservation. VISSA readily shows the four alpha helices of this domain and the alignment is consistent with each protein's predicted secondary structure. The revised seed and complete alignments are shown in Figures B.2 and B.3 of Appendix B, respectively. The representative alignment (Figure 5.2) and its information content (Figure 5.3A) revealed several conserved residues. The most conspicuous ones are two tyrosine residues in the second and third alpha helices that appear to interact in maintaining the bundle as revealed by visualizing the corresponding positions in the Tar structure (Figure 5.3B). The overall conservation pattern of hydrophobic and polar residues reflects the coiled coil nature of this domain. No obvious conserved binding sites can be seen within the alignment, which is expected as known homologs bind very different ligands: aspartate and maltose-binding protein by Tar, serine by Tsr, ribose- and galactose-binding proteins by Trg, and dipeptides by Tap (Falke and Hazelbauer, 2001). Interestingly, the recently described CHASE3 domain family (Zhulin, et al., 2003) was found to comprise a relatively small (approximately 12%) subfamily within 4HB_MCP. The 4HB_MCP containing sensory proteins were found in many major bacterial lineages; however, it is completely missing from eukaryotes and there are only few instances of their presence in Archaea, which are

likely the result of horizontal gene transfer. This fact strongly suggests a bacterial origin for the 4HB_MCP domain.

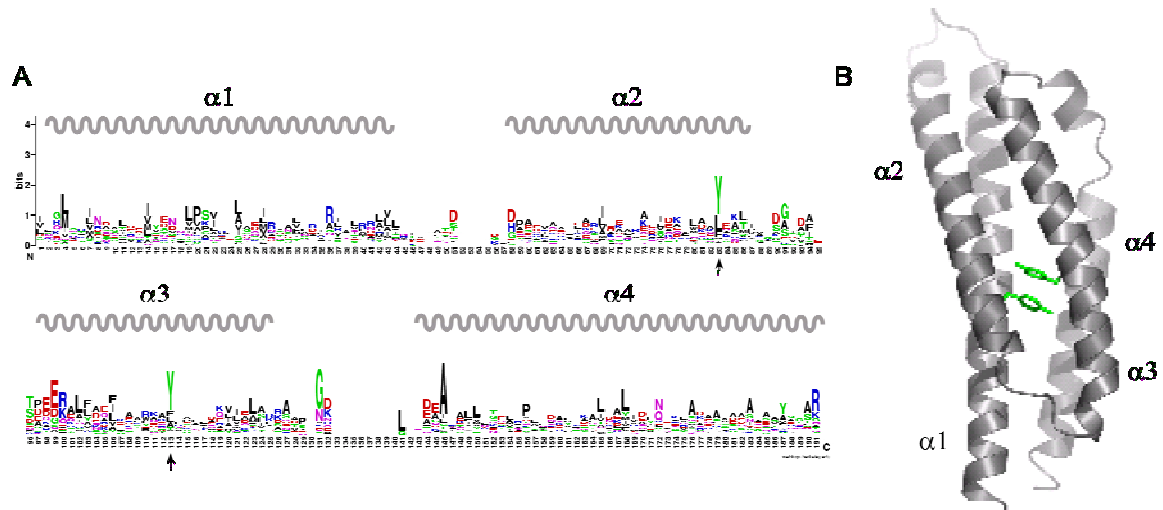


Figure 5.3 Visualization of conserved residues in the 4HB_MCP domain. (A) A sequence logo generated for the multiple alignment of 282 domain sequences obtained in the PSI-BLAST search initiated with the ligand-binding domain of the *E. coli* Tar chemoreceptor. The secondary structure of Tar is shown above the logo. (B) Conserved tyrosine residues in helices 2 and 3 interact to maintain packing of the four-helix bundle.

Domain architectures of all 4HB_MCP-containing proteins were determined using the Pfam and SMART domain libraries (Figure 5.4). The 4HB_MCP domain was found in four major classes of prokaryotic signal transduction: MCPs, sensor histidine kinases, di-guanylate cyclases and di-esterases, and guanylate/adenylate cyclases. Furthermore, this domain is found as an orphan sensory domain and in combination with

other sensory domains as an independent signal transduction module. The 4HB_MCP domain is always found between two predicted transmembrane helices, indicating that it solely detects extracellular signals. In most cases, this domain is associated with a cytoplasmic HAMP domain (Aravind and Ponting, 1999) suggesting that most 4HB_MCP proteins might share the mechanism of transmembrane signaling, which has been extensively studied in *E. coli* chemoreceptors (Ottemann, et al., 1999).

In conclusion, we have identified a large domain superfamily that plays an important role in detecting extracellular signals by bacterial transmembrane receptors involved in various signal transduction pathways. Furthermore, we have developed a visualization technique that uses secondary structure predictions to facilitate producing an accurate alignment of related protein sequences.

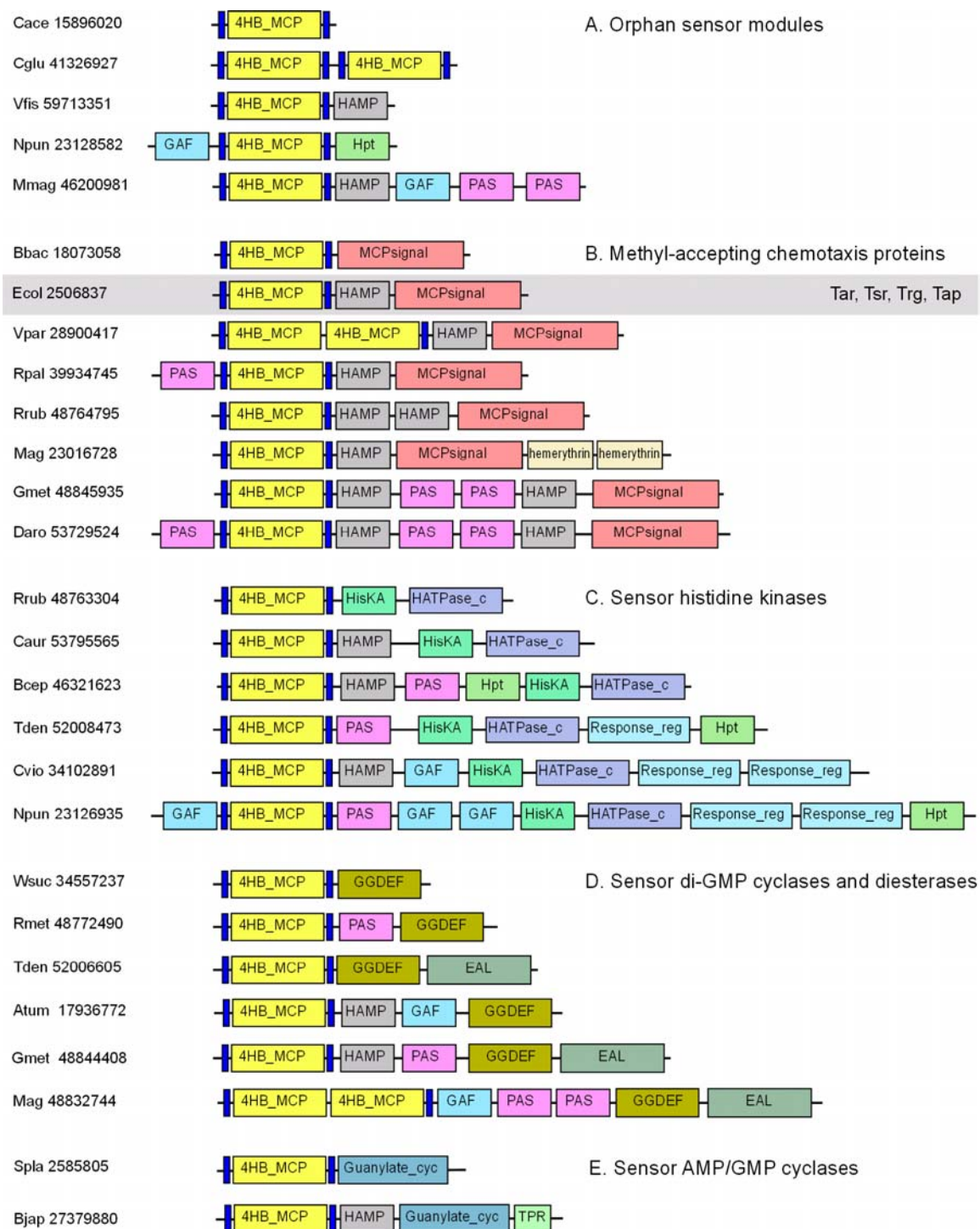


Figure 5.4 A schematic view of the domain architecture of 4HB_MCP containing proteins. Domain nomenclature is according to Pfam (Bateman, et al., 2004). Protein GenBank identifiers and the species abbreviations are shown. Species abbreviations: Atum, *Agrobacterium tumefaciens*; Bbac, *Bdellovibrio bacteriovorus*; Bcep, *Burkholderia cepacia*; Bjap, *Bradyrhizobium japonicum*; Cace, *Clostridium acetobutylicum*; Caur,

Chloroflexus aurantiacus; Cglu, *Corynebacterium glutamicum*; Cvio, *Chromobacterium violaceum*; Daro, *Dechloromonas aromatica*; Ecol, *Escherichia coli*; Gmet, *Geobacter metallireducens*; Mag, *Magnetococcus* sp.; Mmag, *Magnetospirillum magnetotacticum*; Npun, *Nostoc punctiforme*; Rmet, *Ralstonia metallidurans*; Rrub, *Rhodospirillum rubrum*; Spla, *Spirulina platensis*; Tden, *Thiobacillus denitrificans*; Vfis, *Vibrio fischeri*; Vpar, *Vibrio parahaemolyticus*; Wsuc, *Wolinella succinogenes*.

CHAPTER 6

RESOLVING THE FUNCTION OF CHEMOTAXIS PAS DOMAINS THROUGH PROTEIN SEQUENCE ANALYSIS

This chapter is ready for publication. It was recently submitted to the *Bioinformatics* journal and rejected after peer-review; however, the editor recommended addressing the reviewer's concerns and resubmitting the revised manuscript.

Abstract

PAS domains constitute a widespread superfamily of sensory modules. Current computational techniques identify PAS domains in thousands of protein sequences; however they fail to predict the specific function of a given PAS domain. We sought to design a sequence analysis-based approach that could resolve the function of chemotaxis PAS domains.

We identified 274 PAS domains found in association with chemotaxis receptors and classified them into distinct subfamilies. The PAS_Aer subfamily contains the experimentally characterized, FAD-containing redox sensor, Aer, and several critical residues for its function as a redox sensor are solely conserved in this subfamily. We predict that all PAS_Aer domains are redox sensors, whereas members of another subfamily, PAS_Che are predicted to bind a specific, but unknown ligand and represent a good target for experimental studies. Here we demonstrate that standard bioinformatics approaches may be applied to multifunctional domains in order to further resolve their function.

Introduction

Protein domains are structurally compact, independently folding units of a protein. Such structural units have identifiable patterns of amino acid conservation that can be recognized and modeled using profile hidden Markov models (HMMs) (Eddy, 1998). Built from a multiple sequence alignment of related sequences, an HMM putatively represents the structure and function of a protein domain. Many protein domains represented as HMMs are stored in the primary domain databases, Pfam (Bateman, et al., 2004) and SMART (Letunic, et al., 2004). Although HMMs capture the structural properties of a domain, it is often difficult to derive an exact biological function from general profiles that characterize large domain superfamilies. This is especially true for domains involved in signal transduction, as many of them are highly variable in sequence content and length and are associated with a broad spectrum of other domains.

One of the most diverse and functionally important domains employed in signal transduction is the PAS superfamily (Ponting and Aravind, 1997; Zhulin and Taylor, 1997). PAS (Per-ARNT-Sim) domains comprise a widespread superfamily of sensory input modules that sense light, oxygen, redox potential, small ligands, and participate in protein-protein interactions (Zhulin and Taylor, 1997). Structurally, the PAS fold (approximately 100 amino acids in length) consists of four major segments: a loosely-conserved N-terminal cap, the PAS core, a helical linker, and the beta scaffold (Pellequer, et al., 1998). The helical linker connects the PAS core and beta scaffold such that the beta strands form a central sheet surrounded by the remaining helices and loops. Altogether, these structural segments resemble a left-handed glove, which often binds a specific cofactor. In several experimental studies, the specific function of a PAS domain has been

shown to depend upon the type of a bound cofactor. For example, photo-active yellow protein (PYP) contains a 4-hydroxy-cinnamyl chromophore which is responsive to blue light (Genick, et al., 1998), FixL has a heme group coordinated to a histidine residue that is able to directly bind oxygen (Gong, et al., 1998), Aer and NifL proteins are able to detect changes in redox potential via their FAD-containing PAS domains (Macheroux, et al., 1998; Repik, et al., 2000), and plant phototropin is a light sensor utilizing an FMN-containing PAS domain (Harper, et al., 2003).

Until recently, both Pfam and SMART represented the PAS domain as two separate motifs, PAS and PAC, because of significant sequence variation in the helical linker region that was difficult to model; however, these two motifs form a single structural and functional unit and should not be considered as separate domains (Hefti, et al., 2004; Taylor and Zhulin, 1999; Taylor and Zhulin, 1999). The NR database contains thousands of proteins with PAS domains identifiable using the Pfam and SMART domain profiles.

Although PAS domains can be rapidly detected using the Pfam and SMART domain definitions, these profiles only identify the common PAS fold and fail to indicate the specific function of a given PAS domain. We chose to further characterize PAS domains because they represent a significant, extensive, multi-functional superfamily currently represented by a very general domain profile. Furthermore, there is substantial experimental data available to validate and corroborate this analysis including several known structures of PAS domains. In order to constrain the data for this study, analysis was restricted to PAS domains found in association with bacterial methyl-accepting chemotaxis proteins (MCPs). These proteins may be readily identified using the Pfam

and SMART domain profiles. The well-studied Aer transducer of *Escherichia coli* is a representative of this class of signal transduction proteins (Taylor, et al., 1999).

Here we demonstrate that chemotaxis PAS domains can be separated into functionally distinguishable subfamilies using protein sequence analysis. Based on extensive experimental data on the functional role of the PAS domain from the Aer protein, we were able to predict that all related members within this subfamily will bind FAD and sense changes in redox potential. The newly built HMM for this subfamily enables the automatic detection of many important microbial redox sensors in public databases. Additionally, this investigation revealed a distinct subfamily with substantial sequence conservation, which is predicted to bind a specific (although unknown) ligand and represents a good target for experimental studies. Using chemotaxis PAS domains as a case study, we have laid the groundwork for developing a more general approach to functionally resolving existing, multifunctional domains and increasing the accuracy and quality of annotation within public sequence databases.

Methods

HMM searches (Eddy, 1998) seeded with profile hidden Markov models from Pfam version 16.0 and SMART version 3.5, were carried out against the NR database (January 4, 2005). First, we identified 2021 MCP sequences in the NR database using the Pfam MCPsignal (accession PF00015) and SMART MA (accession SM00283) domain profiles. PAS domains in these MCPs were detected with the PAS and PAC domain profiles from both Pfam (accession PF00989 and PF00785) and SMART (accession SM00086 and SM00091). In many cases, only the PAS or the PAC element was detected. Since PAS domains always consist of both the PAS and PAC motifs, the “missing” motif

was delineated from pairwise alignments to other PAS domains from a PSI-BLAST (Altschul, et al., 1997) search with this region. If no significant hits were found, the PAS domain was excluded from the analysis. Current Pfam and SMART profiles fail to recognize all PAS domains. In order to address this problem, we used sensitive PSI-BLAST searches (inclusion threshold $E = 0.01$) initiated with the N-terminal regions of all 2021 MCPs retrieved from the NR-database. Additionally, we included several PAS domains that existed as a single protein yet were predicted to interact with MCPs based on genome neighborhood. These were confirmed experimentally (Hendrixson, et al., 2001). Partial or incomplete PAS domains and duplicate sequences were excluded from this analysis.

A multiple sequence alignment of PAS domains was constructed using the PCMA (Pei, et al., 2003) and ClustalW (Thompson, et al., 1994) programs and manually adjusted (Figure 6.1). The alignment was further tweaked using the VISSA visualization (Ulrich and Zhulin, 2005). Sequence conservation information and the visualized secondary structure of the PAS domains enabled us to produce a much better quality alignment than the initial, unedited alignment constructed by PCMA (Appendix C, Figures C.1 and C.2).

In order to depict the relationship between these PAS domains, we generated a neighbor-joining tree (Saitou and Nei, 1987) of the alignment based on sequence similarity (p-distance) using the MEGA3 software package (Kumar, et al., 2004). Profile HMMs were then built and calibrated for each major cluster within the neighbor-joining tree using the HMMER2 software package (<http://hmmer.wustl.edu>). To increase the sensitivity of each profile, the sequences were weighted based on the Krogh/Mitchison

maximum entropy algorithm (Krogh and Mitchison, 1995). Each specialized PAS HMM was then used in HMMER searches against the NR database to ensure that all members in the respective subfamily were retrieved as well as to identify any new PAS homologs. Strict and relaxed inclusion thresholds were determined and incorporated into the HMM for each PAS subfamily. The strict threshold was set as the bit score of the lowest scoring PAS domain of respective subfamily. The relaxed threshold was set as the bit score of the lowest scoring hit that was distinct from other PAS HMM searches. Searches with these cutoff thresholds revealed that both PAS_Aer and PAS_Che are completely specific and the difference in E-value between the last true positive and first true negative for each subfamily is given in Table 6.1. We evaluated the performance of these new PAS domain profiles and found that they significantly outperform all previous and current PAS profiles from Pfam and SMART (Table 6.2).

Table 6.1 E-values and scores of the last true positive and first true negative from searches of the PAS_Aer and PAS_Che profiles against the non-redundant database (4 January 2005). Scores are given in the parenthesis following the E-value.

<i>Domain profiles</i>	<i>Last true positive</i>	<i>First true negative</i>	<i>Difference</i>
PAS_Aer	2.771e-11 (50.24)	9.534e-11 (44.55)	6.736e-11 (5.684)
PAS_Che	1.707e-20 (86.78)	7.211e-12 (31.39)	7.211e-12 (55.39)

Table 6.2 Identification of PAS domains in methyl-accepting chemotaxis proteins by HMM domain profiles.				
<i>Domain profiles</i>	<i>Number of complete PAS domains identified</i>	<i>Number of incomplete PAS domains identified</i>		<i>Total number of PAS domains identified</i>
		<i>PAS motif only</i>	<i>PAC motif only</i>	
Pfam 17.0 (PAS)	160	N/A	N/A	160
SMART 4.0 (PAS, PAC)	203	35	29	267
This study (PAS_Aer, PAS_Che, generic PAS)	274	N/A	N/A	274

Signal peptides and transmembrane regions with their topology were assigned using Phobius (Kall, et al., 2004). Consensus sequence patterns were generated using the CONSENSUS script by Nigel Brown (<http://www.bork.embl-heidelberg.de/Alignment/consensus.html>). Sequence logos based on the information content were created with the Berkeley weblogo server (Crooks, et al., 2004). A position in the MSA was considered strongly conserved if either the BLOSUM group consensus was greater than 85% or the information content was greater than three bits.

Results and Discussion

Searches with Pfam and SMART domain profiles yielded 240 complete and partial PAS domains associated with MCPs. PSI-BLAST searches revealed an additional twenty PAS domains and enabled the reconstruction of all partially detected PAS domains (e.g. PAC motif only) from domain searches. Finally, fourteen PAS domains comprising individual proteins, but predicted to interact with MCPs on the basis of genome neighborhood (encoded within the same operon), and experimental evidence

(Hendrixson, et al., 2001) were also included in this analysis. We constructed a multiple sequence alignment of these PAS domains and edited it using the VISSA protocol (Ulrich and Zhulin, 2005). The VISSA technique greatly facilitated the process of producing a high-quality alignment by revealing inconsistencies in the MSA such as misalignment of structural elements (Appendix C, Figures C.1 and C.2). Furthermore, VISSA corroborates these results by showing a high degree of agreement between the overall secondary structure of the alignment and the known structure of several PAS domains.

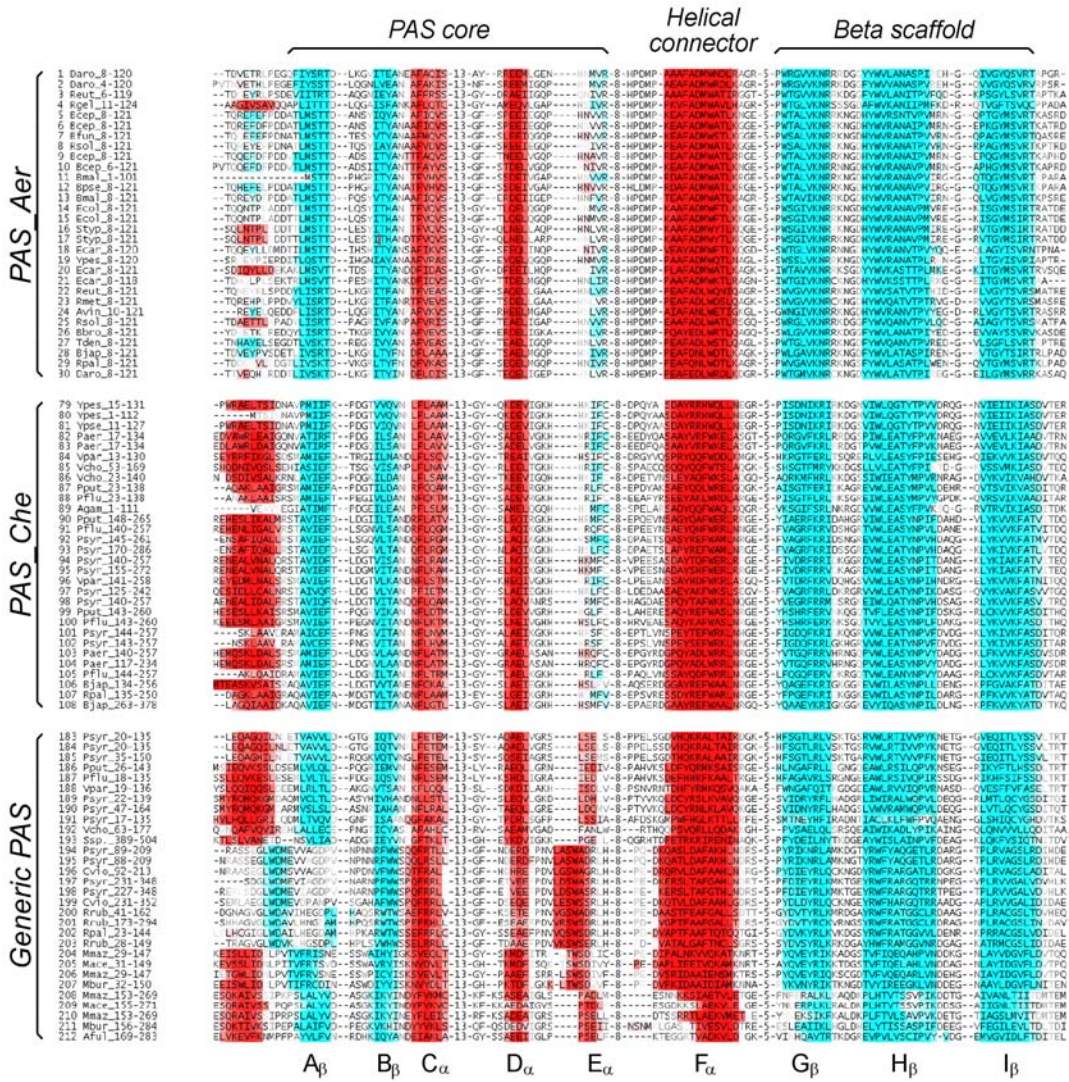


Figure 6.1 Multiple sequence alignment with the VISSA visualization of the three subfamilies of PAS domains found in methyl-accepting chemotaxis proteins. Thirty representative PAS domains from each subfamily are shown (for the full alignment of 274 sequences, see Appendix C, Figure C.2). The multiple alignment was constructed using the PCMA (Pei, et al., 2003) and ClustalW (Thompson, et al., 1994) programs, visualized using the VISSA protocol (Ulrich and Zhulin, 2005). Regions containing predicted alpha helices and beta strands have the background shaded in red and blue, respectively. The shading intensity is directly proportional to the confidence of a given prediction – a darker intensity representing a higher confidence. The PAS core, helical connector, and beta scaffold structural regions are delineated above the alignment. Structural elements, as defined by Gong *et al.* (1998), are listed below the alignment.

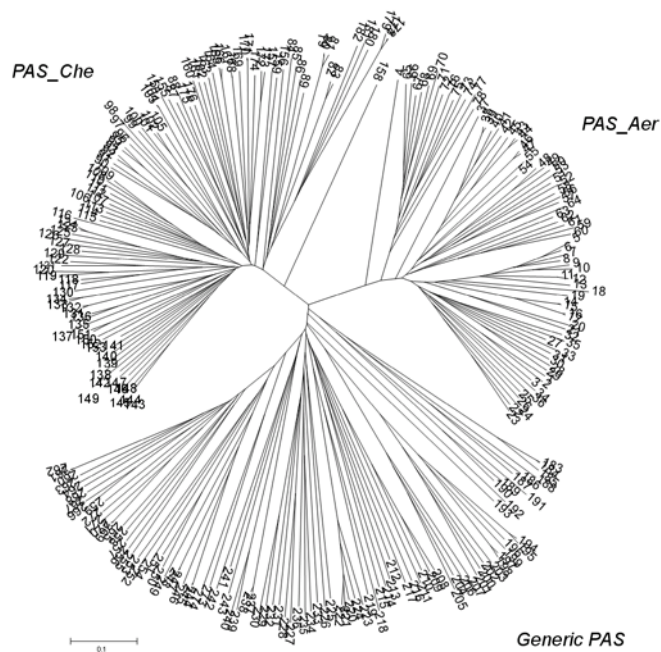


Figure 6.2 Neighbor-joining tree built from a multiple sequence alignment of 274 PAS domains showing three distinct subfamilies.

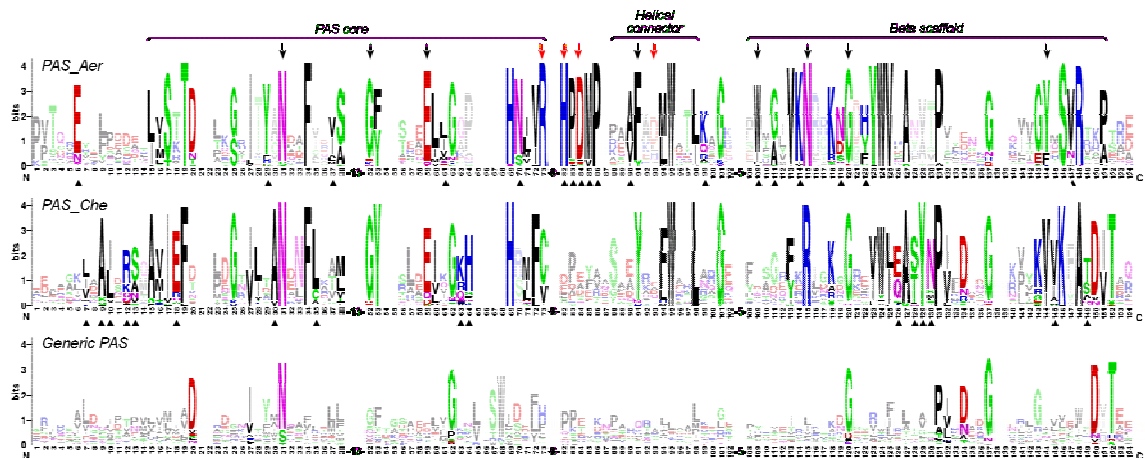


Figure 6.3 Sequence logos for three PAS subfamilies. Highlighted positions are strongly conserved (BLOSUM consensus at least 85% or

information content greater than three bits). Residues unique to a given subfamily are designated with a triangle (▲). Red arrows indicate residues critical for binding FAD by the PAS domain in *E. coli* Aer and black arrows indicate other residues that are critical for Aer function (Repik, et al., 2000).

A neighbor-joining tree built from this high quality alignment revealed two distinct, conserved clusters and one divergent set of PAS domains (Figure 6.2). We designated these three groups as PAS_Aer, PAS_Che, and generic PAS, respectively, based on sequence conservation and the domain architecture of proteins that contain these domains (see below). The secondary structure predictions for each PAS domain are largely consistent with known PAS structures. Six positions (I27, N31, G62, G120, P131, and G137) are strongly conserved across all three groups. Mutational studies indicate that N31 stabilizes the BC turn by forming hydrogen bonds to three backbone nitrogen atoms (Pellequer, et al., 1999; Pellequer, et al., 1998). P131 resides at the C-terminal end of H_β and in conjunction with G137, is responsible for the sharp HI loop of the beta scaffold. Twenty-four other residues are moderately conserved (information content greater than 2 bits) across the three groups (Figure 6.3).

The PAS_Aer subfamily consists of seventy-eight PAS domains including the PAS domain of the experimentally studied *E. coli* aerotaxis transducer, Aer. Fifty-seven amino acid positions are strongly conserved and eighteen of them are unique to this subfamily. All but one of these eighteen positions are conserved in the *E. coli* Aer protein. The most prominent feature that distinguishes this group is the conserved RHPDMP motif located within the EF loop (Figure 6.3). Experimental evidence indicates

that R73, H82, and D84 of this motif (the RHPDMP motif) are crucial for binding the FAD cofactor in Aer (Repik, et al., 2000). The predicted secondary structure of the PAS_Aer subfamily is largely consistent with the overall PAS fold except for the N-terminal cap, which lacks the well-defined alpha helix found in PAS_Che and generic PAS (Figure 6.1). Interestingly, a charged amino acid (predominantly E6) is uniquely conserved within this region indicating its possible functional role (Figure 6.3). Previous experimental studies (Repik, et al., 2000) identified twelve key residues in the PAS domain of Aer that were essential for producing an aerotactic response. Four of these residues (N31, G52, E59, and G120) are conserved throughout the entire alignment and loss of function by cysteine replacement was most likely related to disruption of the PAS fold. Four other residues (R73, H82, D84, and D93) are essential for FAD binding (see above). The remaining four residues (F91, W109, N115, F144) are located within secondary structure elements and show strong conservation only within the PAS_Aer subfamily (Figure 6.3).

Database searches with an HMM built for the PAS_Aer subfamily revealed that this version of the PAS domain is not unique to MCPs and can be found in both one- and two-component signal transduction systems (Ulrich, et al., 2005), such as di-guanylate cyclases and histidine kinases. The domain architecture of PAS_Aer containing MCPs is well conserved and typically includes a single PAS domain followed by one or two transmembrane regions, a HAMP domain, and the C-terminal MCP signaling domain. Occasionally, these (and only these) PAS domains exist as single proteins but are predicted to interact with the corresponding MCPs that contain a HAMP domain, but lack an N-terminal, sensing domain and reside adjacent to the PAS_Aer domains on the

chromosome. In addition to *E. coli* Aer, two other PAS_Aer-containing proteins have been experimentally studied. Both of them, the Aer homolog from *Pseudomonas putida* (Nichols and Harwood, 2000) and the CetA-CetB bipartite (PAS_Aer and the corresponding MCP) system from *Campylobacter jejuni* (Hendrixson, et al., 2001) were shown to govern aerotaxis and related responses in these organisms.

The PAS_Che subfamily comprises 104 PAS domains that share strong sequence similarity. This subfamily contains fifty-eight strongly conserved residue positions and sixteen of them are unique (Figure 6.3). The predicted secondary structure of each of the PAS_Che domains agrees with the known PAS structures. The length of loops/turns in the PAS_Che subfamily is highly conserved (as it is in the PAS_Aer subfamily).

Database searches with the PAS_Che profile confidently identified this type of PAS domains predominantly within methyl-accepting chemotaxis proteins (hence, the name PAS_Che); however, members of the PAS_Che domain subfamily can also be found in two other major classes of prokaryotic signal transduction, namely, histidine kinases and di-guanylate cyclases. MCPs containing PAS_Che domains have one to four copies of PAS_Che and occasionally one or two generic PAS domains. The prominent feature of this subfamily is that PAS_Che-containing MCPs lack any predicted transmembrane regions and therefore are likely to be exclusively cytoplasmic receptors. Only one PAS_Che-containing protein, the McpY of *Sinorhizobium meliloti* was experimentally studied and found not to be an aerotaxis/energy taxis transducer (Scharf, Schmitt, Zhulin and Taylor, unpublished data). Furthermore, although McpY contributes to the overall chemotactic response, the reconstituted McpY protein does not bind FAD (Meier,

Muschler and Scharf, unpublished data). Thus, the intracellular signal recognized by this MCP remains unknown.

The third group of PAS domains found in MCPs consists of ninety-two members and both the alignment and neighbor-joining tree reveal a highly divergent set of sequences. The predicted secondary structure still follows that of a typical PAS fold, although several sequences contain an extension of the C $_{\alpha}$ helix and appear to be missing the E $_{\alpha}$ helix. Thirty-five PAS domains in this subset have a reduced helical connector and a shorter G $_{\beta}$ strand. Only eleven residues are conserved in generic PAS domains and none of these positions are unique to this subset. Such remarkable lack of conservation is in striking contrast to the PAS_Aer and PAS_Che subfamilies and suggests that no specific function is associated with this group (hence the name, “generic PAS”). Searches with the profile HMM built for this group of sequences recognize PAS domains in numerous proteins representing all major classes of prokaryotic signal transduction. In MCPs, generic PAS domains occur in a single protein in up to five copies and can co-occur with either PAS_Aer or PAS_Che domains.

No MCPs that contain generic PAS domains have been studied experimentally and elucidation of their function, as well as the function of many other PAS domain-containing signal transduction proteins, necessitates high-quality, high-throughput sequence analysis of the entire PAS superfamily.

In summary, the high-quality, microbial signal transduction data derived by these analytical techniques offer promising new perspectives and expectations from the available, underlying genetic information.

Acknowledgements

We extend our thanks to our co-authors: William Black, Qinhong Ma, and Barry Taylor. We also thank Meier, Muschler, Scharf, Schmitt, Zhulin, and Taylor for their supportive role in providing unpublished results.

CHAPTER 7

CONCLUSION

In conclusion, this research significantly contributes to the understanding of microbial signal transduction. A bioinformatics platform and the Microbial Signal Transduction database (MiST) were developed that enabled the comprehensive, integrated analysis of microbial signal transduction at the domain level. This platform was used successfully for the high-throughput identification and classification of signal transduction systems in more than 300 bacterial and archaeal organisms. Using this infrastructure, we performed a comprehensive review of signal transduction systems and found that contrary to the current view, the majority of signal transduction systems consist of one-component systems – a single protein containing input and output domains but lacking phosphotransfer domains typical of two-component systems. Two-component systems utilize a subset of the input and output domains found in one-component systems, indicating a reduced repertoire of sensory and response modules. Furthermore, one-component systems display extraordinary combinatorial domain diversity. The dominance of one-component systems in prokaryotic signal transduction is a paradigm-shifting discovery rendering a more accurate depiction of the scope and complexity of bacterial signal transduction. A novel source of information regarding signal transduction systems, the MiST database has been effectively used for annotating novel genomes and performing indepth characterization of sensory domains. Altogether, this systematic, high-throughput delineation of microbial signal transduction is another step forward in our understanding of the genomic basis of life.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

Table A.1 Domains and domain categories used to identify signal transduction systems.

Input domains						
<i>Small-molecule binding</i>	<i>Enzymatic</i>	<i>Cofactor binding</i>	<i>Unknown function</i>	<i>Protein-protein interactions</i>		
ACT	aminotran_1_2	BLUF	CHASE2	CBS		
Ada_Zn_binding	Archaeal_ATPase	fer4	CHASE3	TPR		
AlkA_N	citrate_synt	FeS	CHASE4			
AraC_binding	Cyanate_lyase	Hemerythrin	MASE1			
Autoind_binding	EPSP_synthase	HhH-GPD	MASE2			
Cache	FmddA_AmdA	NIR_SIR	MHYT			
CHASE	GATase_2	NIR_SIR_ferr	TrkA-C			
cNMP_binding	Glucokinase	Nitro_FeMo-Co				
Diacid_rec	Glycos_trans_3N	phytochrome				
Fe_dep_repr_C	Glyoxalase					
FeoA	HEAT_PBS					
FHA	HEM4					
GAF	Nitroreductase					
HMA	NTP_transf_2					
LysR_substrate	NUDIX					
PAS	PALP					
Peripla_BP_2	Peptidase_M37					
Peripla_BP_like	peroxidase					
SBP_bac_3	pfkB					
SIS	Pribosyltran					
STAS	PTS_EIIC					
TOBE	PTS-HPr					
V4R	pyr_redox					
NIT*	Rhodanese					
	SKI					
Output domains						
<i>DNA-binding</i>	<i>Di-guanylate cyclase</i>	<i>RNA-binding</i>	<i>Phosphatase</i>	<i>Protein kinase</i>	<i>Hydrolase</i>	<i>Other</i>
Arc	EAL	ANTAR	PP2C_SIG**	pkinase	HD	guanylate_cyc
Arg_repressor	GGDEF	CsrA				LytR_cpsA_psr
ASNC_trans_reg						Rrf2
crp						RseA_N
CtsR						
deoR						
Fe_dep_repress						
GerE						
gntR						

Table A.1 (continued)

HTH_1
 HTH_3
 HTH_4
 HTH_5
 HTH_6
 HTH_7
 HTH_8
 HTH_9
 HTH_10
 HTH_AraC
 IclR
 AlcI
 LytTR
 MarR
 MerR
 PadR
 ROS_MUCR
 TetR
 trans_reg_C

Transmitter / receiver domains

<i>Histidine kinases</i>	<i>Response regulator</i>	<i>Other</i>
HATPase_c	response_reg	HAMP
HisKA		MCPsignal
		Sigma54_activat

All domain nomenclature and assignments are according to Pfam (ref)

* Domain is from (ref)

** Domain is from the SMART database (ref)

Table A.2 Distribution of two-component and one-component systems in prokaryotic genomes.

	Genome size (Mb)	Two-component systems		One-component systems
		Histidine kinases	Response regulators	
Archaea				
Crenarchaeota				
<i>Aeropyrum pernix</i>	1.7	0	0	14
<i>Pyrobaculum aerophilum</i>	2.2	0	0	25
<i>Sulfolobus solfataricus</i>	3	0	0	49
<i>Sulfolobus tokodaii</i>	2.7	0	0	51
Euryarchaeota				
<i>Archaeoglobus fulgidus</i>	2.2	14	11	49

Table A.2 (continued)

<i>Methanosarcina acetivorans</i>	5.8	53	15	66
<i>Methanosarcina mazei</i>	4.1	34	14	50
<i>Halobacterium</i> sp.	2.6	13	5	56
<i>Methanobacterium thermoautotrophicum</i>	1.8	16	7	24
<i>Methanococcus jannaschii</i>	1.7	1	0	29
<i>Methanosarcina barkeri</i>	5.1	23	10	47
<i>Methanopyrus kandleri</i>	1.7	0	0	17
<i>Pyrococcus abyssi</i>	1.8	1	2	32
<i>Pyrococcus furiosus</i>	1.9	0	0	30
<i>Pyrococcus horikoshii</i>	1.7	1	2	30
<i>Thermoplasma acidophilum</i>	1.6	0	0	22
<i>Thermoplasma volcanium</i>	1.6	0	0	21
<i>Ferroplasma acidarmanus</i>	1.9	0	0	32
Bacteria				
<u>Aquificae</u>				
<i>Aquifex aeolicus</i>	1.6	3	5	34
<u>Thermotogae</u>				
<i>Thermotoga maritima</i>	1.9	10	11	57
<u>Planctomycetes</u>				
<i>Pirellula</i> sp.	7.1	45	56	137
<u>Cyanobacteria</u>				
<i>Synechococcus</i> sp.	2.4	6	9	16
<i>Synechocystis</i> PCC6803	3.6	42	40	68
<i>Thermosynechococcus elongatus</i>	2.6	19	23	40
<i>Nostoc</i> sp.	7.2	137	80	148
<i>Nostoc punctiforme</i>	9	164	100	146
<i>Trichodesmium erythraeum</i>	7.7	35	28	69
<i>Prochlorococcus marinus</i>	1.8	4	6	9
<u>Deinococcus</u>				
<i>Deinococcus radiodurans</i>	3.3	20	24	102
<u>Actinobacteria</u>				
<i>Bifidobacterium longum</i>	2.3	10	9	67
<i>Corynebacterium efficiens</i>	3.1	13	13	91
<i>Corynebacterium glutamicum</i>	3.3	13	13	111
<i>Mycobacterium bovis</i>	4.3	15	12	162
<i>Mycobacterium leprae</i>	3.3	5	5	38
<i>Mycobacterium tuberculosis</i>	4.4	15	12	159
<i>Streptomyces avermitilis</i>	9	132	73	519
<i>Streptomyces coelicolor</i>	9.1	143	85	627
<i>Tropheryma whipplei</i>	0.9	2	2	7
<i>Thermobifida fusca</i>	3.6	40	26	159
<u>Firmicutes</u>				
<i>Bacillus anthracis</i> Ames	5.2	49	47	230

Table A.2 (continued)

<i>Bacillus cereus</i>	5.4	55	46	224
<i>Bacillus halodurans</i>	4.2	46	48	174
<i>Bacillus subtilis</i>	4.2	39	35	187
<i>Listeria innocua</i>	3.1	18	18	150
<i>Listeria monocytogenes</i>	2.9	17	17	147
<i>Oceanobacillus iheyensis</i>	3.6	22	21	145
<i>Staphylococcus aureus</i>	2.9	18	17	72
<i>Staphylococcus epidermidis</i>	2.5	17	16	50
<i>Clostridium acetobutylicum</i>	4.1	38	43	174
<i>Clostridium perfringens</i>	3.1	28	20	95
<i>Clostridium tetani</i>	2.8	31	28	88
<i>Thermoanaerobacter tengcongensis</i>	2.7	20	21	79
<i>Clostridium thermocellum</i>	3.7	32	32	75
<i>Desulfitobacterium hafniense</i>	7	67	81	273
<i>Enterococcus faecalis</i>	3.2	15	18	138
<i>Lactobacillus plantarum</i>	3.3	11	15	189
<i>Lactococcus lactis</i>	2.4	8	7	96
<i>Streptococcus agalactiae</i>	2.2	18	21	67
<i>Streptococcus mutans</i>	2	13	14	90
<i>Streptococcus pneumoniae</i>	2.2	13	14	61
<i>Streptococcus pyogenes</i>	1.9	10	12	67
<i>Enterococcus faecium</i>	3	13	16	92
<i>Lactobacillus gasseri</i>	2	5	5	65
<i>Leuconostoc mesenteroides</i>	2.1	7	7	77
<i>Oenococcus oeni</i>	1.9	6	6	63
<i>Mycoplasma gallisepticum</i>	1	0	0	2
<i>Mycoplasma genitalium</i>	0.6	0	0	4
<i>Mycoplasma penetrans</i>	1.4	0	0	9
<i>Mycoplasma pneumoniae</i>	0.8	0	0	3
<i>Mycoplasma pulmonis</i>	1	0	0	1
<i>Ureaplasma urealyticum</i>	0.8	0	0	3
Fusobacteria				
<i>Fusobacterium nucleatum</i>	2.2	8	7	37
<i>Fusobacterium nucleatum subsp. vincentii</i>	2.1	6	4	31
Chlamydiae				
<i>Chlamydia muridarum</i>	1.1	2	1	6
<i>Chlamydia trachomatis</i>	1	2	1	6
<i>Chlamydophila caviae</i>	1.2	2	1	8
<i>Chlamydophila pneumoniae</i>	1.2	2	1	7
Spirochaetes				
<i>Borrelia burgdorferi</i>	1.5	5	6	3
<i>Leptospira interrogans</i>	4.7	48	32	101
<i>Treponema pallidum</i>	1.1	3	3	13
Bacteroidetes				
<i>Bacteroides thetaiotaomicron</i>	6.3	76	30	98

Table A.2 (continued)

<i>Porphyromonas gingivalis</i>	2.3	6	6	24
<i>Cytophaga hutchinsonii</i>	4.4	56	32	60
<u>Chlorobi</u>				
<i>Chlorobium tepidum</i>	2.2	9	4	17
<u>Chloroflexi</u>				
<i>Chloroflexus aurantiacus</i>	4.9	70	66	106
<u>Proteobacteria - epsilon</u>				
<i>Campylobacter jejuni</i>	1.6	7	11	9
<i>Helicobacter hepaticus</i>	1.8	5	10	10
<i>Helicobacter pylori</i>	1.7	3	9	5
<i>Wolinella succinogenes</i>	2.1	38	43	53
<u>Proteobacteria - delta</u>				
<i>Desulfovibrio desulfuricans</i>	3.9	51	66	112
<i>Geobacter metallireducens</i>	4.2	88	74	101
<u>Proteobacteria - magnetotactic</u>				
<i>Magnetococcus sp.</i>	4.7	95	73	108
<u>Proteobacteria - alpha</u>				
<i>Magnetospirillum magnetotacticum</i>	9.2	126	144	246
<i>Novosphingobium aromaticivorans</i>	4.4	27	33	156
<i>Rhodobacter sphaeroides</i>	4.6	41	48	170
<i>Rhodopseudomonas palustris</i>	5.5	55	55	227
<i>Rhodospirillum rubrum</i>	4.7	53	55	186
<i>Agrobacterium tumefaciens</i>	5.7	46	55	353
<i>Bradyrhizobium japonicum</i>	9.1	85	93	446
<i>Brucella melitensis</i>	3.3	20	22	133
<i>Brucella suis</i>	3.3	19	21	131
<i>Caulobacter crescentus</i>	4	53	46	151
<i>Mesorhizobium loti</i>	7.6	52	56	461
<i>Rickettsia conorii</i>	1.3	4	4	8
<i>Rickettsia prowazekii</i>	1.1	4	4	4
<i>Sinorhizobium meliloti</i>	6.7	40	56	390
<i>Rickettsia sibirica</i>	1.3	4	4	8
<u>Proteobacteria - beta</u>				
<i>Burkholderia fungorum</i>	9.6	70	73	545
<i>Ralstonia metallidurans</i>	6.8	61	84	377
<i>Bordetella bronchiseptica</i>	5.3	26	31	382
<i>Bordetella parapertussis</i>	4.8	22	28	312
<i>Bordetella pertussis</i>	4.1	19	21	221
<i>Chromobacterium violaceum</i>	4.8	47	64	233
<i>Neisseria meningitidis</i>	2.3	4	4	30
<i>Nitrosomonas europaea</i>	2.8	15	18	51
<i>Ralstonia solanacearum</i>	5.8	45	61	277
<u>Proteobacteria - gamma</u>				

Table A.2 (continued)

<i>Azotobacter vinelandii</i>	5.4	40	37	225
<i>Microbulbifer degradans</i>	5.4	62	72	156
<i>Blochmannia floridanus</i>	0.7	0	0	5
<i>Buchnera aphidicola</i>	0.6	0	0	1
<i>Buchnera sp.</i>	0.7	0	0	2
<i>Coxiella burnetii</i>	2	6	8	18
<i>Escherichia coli</i>	4.6	30	32	230
<i>Haemophilus ducreyi</i>	1.7	1	3	24
<i>Haemophilus influenzae</i>	1.8	3	6	42
<i>Pasteurella multocida</i>	2.3	9	9	41
<i>Pseudomonas aeruginosa</i>	6.3	64	71	394
<i>Pseudomonas putida</i>	6.2	68	73	350
<i>Pseudomonas syringae</i>	6.4	67	69	283
<i>Salmonella typhi</i>	5.1	30	34	236
<i>Salmonella typhimurium</i>	5	32	35	245
<i>Shewanella oneidensis</i>	5.1	46	57	204
<i>Shigella flexneri</i>	4.6	25	29	166
<i>Vibrio cholerae</i>	4	43	49	195
<i>Vibrio parahaemolyticus</i>	5.2	50	55	287
<i>Vibrio vulnificus</i>	5.1	52	58	283
<i>Wigglesworthia brevipalpis</i>	0.7	0	1	3
<i>Xanthomonas campestris</i>	5.1	55	54	163
<i>Xanthomonas citri</i>	5.2	61	59	167
<i>Xylella fastidiosa</i>	2.7	13	17	40
<i>Xylella fastidiosa Temecula I</i>	2.5	13	18	28
<i>Yersinia pestis</i>	4.8	21	27	198
<i>Actinobacillus pleuropneumoniae</i>	2.3	3	2	28
<i>Haemophilus somnus</i>	2.2	2	2	34
<i>Pseudomonas fluorescens</i>	6.5	75	84	362

Table A.3 Genomic distribution of input and output domains in archaea and bacteria.

Input domains*	Two-component systems		One-component systems	
	Archaea	Bacteria	Archaea	Bacteria
LysR_substrate	0	0	3	2542
PAS	104	485	1	470
GAF	21	284	0	317
Peripla_BP_like	0	3	0	537
AraC_binding	0	0	0	212
cNMP_binding	0	7	0	177
Cache	4	109	0	5
SIS	0	0	1	117

Table A.3 (continued)

aminotran_1_2	0	0	0	76
TPR	0	7	1	64
CBS	0	10	11	42
Fe_dep_repr_C	0	0	18	43
MHYT	0	3	0	54
SBP_bac_3	0	37	0	20
CHASE3	0	47	0	6
ACT	0	0	0	52
CHASE2	0	12	0	37
Ada_Zn_binding	0	0	0	46
FHA	0	2	0	39
NTP_transf_2	0	0	0	38
NIT	0	32	0	5
CHASE	0	16	0	21
Autoind_bind	0	0	0	35
CHASE4	10	0	0	20
TOBE	0	0	2	28

Output domains*	Two-component systems		One-component systems	
	Archaea	Bacteria	Archaea	Bacteria
HTH_1	0	0	5	2673
tetR	0	0	33	1634
gntR	0	0	6	1383
trans_reg_C	0	1202	0	176
HTH_AraC	0	85	0	1287
GGDEF	0	132	0	1221
HTH_3	0	0	104	1237
GerE	0	618	2	376
MarR	0	0	36	800
EAL	0	71	0	727
HD	0	53	79	570
HTH_8	0	263	0	334
HTH_5	0	0	98	461
lacI	0	0	0	537
merR	0	0	4	493
ASNC_trans_reg	0	0	93	399
pkinase	0	3	13	466
IcIR	0	0	1	379
deoR	0	0	0	285
PP2C_SIG	0	58	1	200
PadR	0	0	39	187
Rrf2	0	0	0	222
guanylate_cyc	0	12	0	185
crp	0	1	1	194
LytTR	0	109	0	51

* Twenty-five most abundant domains are listed

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Ecol_2506837_34-190	-----HHSQKSEFVSNQLRQEQGELTSTWDLMLQTRINLSRSVAVRM-----MDSSQQ-----SNARKV-----SLDSARKV-----
Ecar_50120625_27-192	-----GLFFSALKND-----KDFSSQVINQKRSELDASWYLLQTRNTLNR-----AGI-----RF-----ALDVSGTGA-----VGGKEL-----LSAEKQLAV-----
Styp_16423100_27-192	-----GLFFNSLKND-----KENFTVLOTIRQOQSALNATWVLLQTRNTLNR-----AGI-----RW-----VMDQSNIGSGATVAELN-----OGATNTLKI-----
Sent_56416317_27-192	-----GLFFNSLKND-----KENFTVLOTIRQOQSALNATWVLLQTRNTLNR-----AGI-----RW-----VMDQSNIGSGATVAELN-----OGATNTLKI-----
Ecol_43218_27-192	-----GLFFNALKND-----KENFTVLOTIRQOOSTLNGSWVALLQTRNTLNR-----AGI-----RY-----VMDQNNIGSGSTVAELN-----SSASISLKG-----
Ecol_16132176_27-192	-----GLFFNALKND-----KENFTVLOTIRQOOSTLNGSWVALLQTRNTLNR-----AGI-----RY-----VMDQNNIGSGSTVAELN-----SSASISLKG-----
Ecol_26251236_27-192	-----GLFFNALKND-----KENFTVLOTIRQOOSTLNGSWVALLQTRNTLNR-----AGI-----RY-----VMDQNNIGSGSTVAELN-----SSASISLKG-----
Sfile_24115585_27-192	-----GLFFNALKND-----KENFTVLOTIRQOOSTLNGSWVALLQTRNTLNR-----AGI-----RY-----VMDQNNIGSGSTVAELN-----SSASISLKG-----
Ecol_13364793_27-192	-----GLFFNALKND-----KENFTVLOTIRQOOSTLNGSWVALLQTRNTLNR-----AGI-----RY-----VMDQNNIGSGSTVAELN-----SSASISLKG-----
Bjap_27376721_106-259	-----DCVRY-----QTVYRHMQLDSASAVNIGRNVNLIYAIVMESI-----GIV-----M-----STE-----AK-----VKQFA-----DELVKCSGE-----
Cace_15896003_35-190	-----LEKV-----SGSATKMYKVNLLKKYYILGEMEQLRLEIRADVI-----LIV-----Y-----ORD-----SN-----LKGFA-----FEIVNDEKE-----
Map_48832885_28-180	-----NHLTE-----QSAVDVLERENYRSVIASQEMKEALERMDSGVLE-----FHH-----AGRGHEG-----AAILA-----AGIADFRKA-----
Gmet_48846841_24-186	-----AGLGVKA-----DKASNELLSQEVKIGEYFSRVNANILYMMMYE-----DAE-----IN-----INN-----DK-----LAEYE-----AKWTEKKGR-----
Bbac_42522983_1-181	MIAAAMPVIGFGIVYALSQGMKG-----DAYLENSHKNTIPGQALAEKQARKKYGQYV-----AM-----SV-----KTE-----KR-----DEEL-----AKAEAVKE-----
Bbac_18073058_46-208	-----NGITSV-----VNLLDVANEV-----PTFDMVGEKQARKKFGYQAM-----AM-----E-----IDE-----EK-----LTSYL-----KATATGMDK-----
Bbac_42522982_46-208	-----NGITSV-----VNLLDVANEV-----PTFDMVGEKQARKKFGYQAM-----AM-----E-----IDE-----EK-----LTSYL-----KATATGMDK-----
lloi_56461443_33-194	-----VNFKSY-----ANSVDEITASEHLPGLNLLQADRDHLQAOVAER-----NIL-----SV-----PSGS-----PK-----QOQLT-----DTFNENLQO-----
Mdeg_48864484_35-195	-----LSEI-----SRSSAQVTENNLPATEFLLEADRDHLQSLTAE-----ELV-----LS-----PANC-----VN-----KCAN-----KTLNENLQO-----
Sent_62180191_48-202	-----QADRDQRT-----VTABEITRTGLANSDFLRSARINMIQAGAASR-----IA-----EMD-----KCAN-----AAETRTKG-----
Sfile_24113141_45-198	-----QADRDQRT-----VTABEITRTGLANSDFLRSARINMIQAGAASR-----IA-----EMD-----KCAN-----AAETRTKG-----
Ecol_16129380_45-198	-----QADRDQRT-----VTABEITRTGLANSDFLRSARINMIQAGAASR-----IA-----EMD-----KCAN-----AAETRTKG-----
Ecol_12515286_45-198	-----QADRDQRT-----VTABEITRTGLANSDFLRSARINMIQAGAASR-----IA-----EMD-----KCAN-----AAETRTKG-----
Gmet_48847023_33-190	-----TCLETEM-----KSHLDAMDHDIPG-----ETAAMIDRHHQOHRRTFI-----NYL-----E-----NRLHKELE-----LKLHHEATD-----
Zmob_56542672_36-189	-----RVQWKE-----ASIAKDLGVAQGRARAQVSVDISRAADYRSAT-----RYV-----I-----VPD-----CA-----LQAAE-----QMTVYRKG-----
Npun_53688984_25-201	-----SGSSRI-----SKHIDTLANNVPSISLWKNINAGQTQIESSE-----GIL-----DWNLSKDG-----RQTEI-----DRMDNAWKO-----
Lint_45657908_29-179	-----VDS-----NDRLKRIVDVAKK-----NLSEHILGVLEARHEK-----NII-----I-----EKDP-----K-----MVYR-----DRIYKAVDS-----
Pres_27228663_28-198	-----LPMAGLNGMRTS-----NAVIDLRLYNMLPLDKLGDINNHHMNAQAQL-----LALQHEFGSAFE-----KHD-----HF-----SSMHF-----DRVDKSVSI-----
Dhaf_23120317_31-189	-----IIEIQHRA-----RGALQSLQ-----DSLRGLRLIKITVDSDAYGLDVDDTT-----FRV-----RIDL-----VG-----NDEGV-----DRVDRARAR-----
Mmag_23013876_33-194	-----QGLAGC-----SRTLDEIAEVSFAKSVKIGDFTLLQDAHSDLY-----RLI-----TWNAGV-----BAKTOKTE-----ASFAEGMAK-----
Psyr_28868215_43-205	-----AFCLLMQCEI-----RNQSETVESGALPSIADADAIAIGLVKLRSSET-----RLI-----AN-----ADD-----GS-----VINSE-----INVEQLRNE-----
Psyr_46188215_30-188	-----LQMQAI-----RTQGEAVESGALPSIADADAIAIGLVKLRSSET-----RLI-----AN-----ADD-----GA-----VINSE-----INVEQLRNE-----
Psyr_28870455_33-190	-----DLSSI-----KAKGLEIENDSLPGIALGDAIALAFSNTDYDM-----KML-----SAR-----SAD-----VPOAE-----BELMORNE-----
Psyr_46187691_15-170	-----QLHGI-----RQOSLEIENDALPGIALGDDIALAFKTRTTVA-----KML-----AAH-----DMA-----VPLAH-----EFLDOKAK-----
Psyr_28870175_31-187	-----DQLHGI-----RQOSLEIENDSLPGIALGDDIALAFKTRTTVA-----KML-----LQ-----DVAC-----VALVA-----DPLEKKAG-----
Caur_53795565_28-187	-----BHQMAVN-----NBRRAVTERHTIPSLTVGIMTAALNRYARQL-----EFL-----I-----YT-----N-----G-----RARLI-----QMRVIEH-----
Rgel_47574589_32-189	-----SRIVMV-----ERGFDSALDVER-----VARADEMGLTTLNVSRTI-----LIV-----K-----SGGL-----POM-----SAFPI-----QNKDTSFO-----
Tden_52007873_23-184	-----VIGWRLQTV-----GMDTSLVKNEMKKARIINENWESINANAVRAI-----AAA-----K-----TNN-----ET-----EKFTV-----DASAAASKG-----
Reut_53761194_30-189	-----LYQINQV-----SSSTQVQKQPLRKERLASDWHATLVACVQRM-----AVA-----RS-----NDS-----JVELFAENTRASKE-----
Rmet_48772113_31-189	-----VRLQCV-----AERTHDMMQPLTKERLVSDMYRLMHTSVRRTI-----AVA-----K-----ADD-----SL-----GAFEA-----KATKASLEG-----
Ecar_50123040_34-191	-----KLEDA-----AGRTHAMQV-----LAKERIVSDMYRLMHTSVRRTI-----AIT-----E-----SDF-----SL-----GQFA-----KATKASLEG-----
Ecar_50123255_33-185	-----VGE-----GNDIKRLSGITTLANMLLITQDKAGFDANARLTH-----ATA-----I-----STD-----AG-----IKQBR-----QFVDSQIAR-----
Ecar_50123254_32-191	-----ERL-----SGNQLLLSQIRINLLMQEYKONVNDTARATH-----WMA-----L-----ND-----TQ-----KMTBK-----ERIESAR-----
Rgel_47571710_55-210	-----LQLDKI-----SENQVLSQVRI-----NILLMQEYKONVNDTARATH-----WMA-----L-----ND-----TQ-----KMTBK-----ERIESAR-----
Rgel_47573506_34-195	-----LALQRT-----SDAVDIRVQGEWVKAAGASIDTLTRANARHT-----ELP-----V-----ECS-----AAAVR-----ERIANROG-----
Rgel_47571655_31-195	-----LASE-----HQSETHYVDQVATVRIITLMDVDAANARATISAR-----NLV-----L-----VSA-----AD-----REIBK-----AAVTAHKKV-----
Rsol_17428912_49-203	-----LSSLSGI-----SHQRFVFDVGSQREALANQVMDAAQRAITAA-----NLV-----L-----VTS-----QD-----REIBK-----AAVTAHKKV-----
Reut_46131863_29-194	-----DGR-----NARTIGFVDGVNARVLMVNRIRAAVDRRAIAAR-----NLV-----L-----ATTA-----QD-----RAFEK-----AEAEARHEE-----
Rgel_47572127_31-190	-----LAVKSLND-----NDRFASFVQGENAQLLSLEYVTAIERRAILAR-----DLV-----N-----AAT-----QD-----RATIK-----AEVTVHDD-----
Bcep_46324344_24-189	-----SE-----HDQFVDYHGVNARAAQAAADVRAAVDRRAIAAR-----NLV-----L-----VRK-----AD-----LAABC-----NSVVAARKE-----
Reut_53761951_33-194	-----LAINALNE-----NNRFSGFV-----GINGRAQMAASVRTAVDDRMAVAV-----NLV-----L-----VTA-----AD-----VEIEK-----NSVVAARKE-----
Bcep_46319966_34-197	-----ALGES-----TQGTQYINGLGNARSELSEAEVRAAVDRRAIAAR-----NMV-----L-----ATVP-----AE-----LQKEX-----AEALRAHEE-----
Bmal_53723395_36-194	-----HJ-----NAEFSRYMNGINARATLSAQIRTAVDRRAIAAR-----NLV-----L-----VTK-----SD-----LELEL-----AEVNGAHKI-----
Bfun_48780518_29-192	-----LVSLSA-----HGRFSYDYNGINARATLADRVHVAVDRRAIAAR-----NLV-----L-----AGSA-----AD-----VEHEK-----BAALQADK-----
Rmet_48770099_38-194	-----HA-----NDRFASYVSI-----ISARAAAEQVRTAVDRRAIAAR-----NLV-----L-----VTK-----AD-----VELEK-----AAVTAQBDI-----
Cvio_34102638_19-177	-----KSLGDS-----TDGFTTFVHGVNARADVMVQVRTAVDRRAIAAR-----NLV-----L-----VTK-----QD-----LEIEK-----ADVLRAHEE-----
Cvio_34105172_32-187	-----AYNKLOTI-----EDGFSAYINGLGNARSEMASHVTRVDRRAIAAR-----NLV-----L-----VTE-----AD-----LSELE-----BAVLKAHDD-----
Cvio_34101405_32-187	-----KMDNI-----QNNITEIKEDRYPKVMLLDRIITMTLLDIGRGR-----NAI-----I-----APD-----QD-----VEQOI-----RNVEVLRAK-----
Ragu_15077504_33-190	-----NLSAM-----QAKTRDITEDRYPKVMLLDRIITMTLLDIGRGR-----NAI-----I-----APD-----QD-----VEQOI-----RNVEVLRAK-----
RspH_46192461_24-187	-----YTKISSI-----QANRRBITNNRYPKAKCNKIVITTOEYVSKLIR-----DAV-----L-----SDH-----QD-----LESNI-----SKVVALRAE-----
Daro_53729525_21-178	-----LMLRLGTH-----NSSIEQIISDRYKVKLAFDPVDRGVNDQIKYLE-----GIV-----L-----TKNE-----EQ-----NNKRY-----LQLDSDVKO-----
Bpse_53722895_48-204	-----LNTI-----QOKVDHIVDNNVKKLSLAVDMRNNLIARHVR-----KAL-----I-----YSC-----EK-----QKSEA-----KVPDAFAK-----
Bcep_46316835_48-204	-----LRLRAI-----NDEAISIER-----SLPGYVLAASLRASANEYSIYLO-----RAV-----TV-----DAEA-----EA-----VQDOL-----AKVGLGLEE-----
Bcep_46323416_48-204	-----LRLRAI-----NSEAVSIEDSLPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Rmet_48771949_47-208	-----LRLRAI-----NAEAVSIEDSLPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Reut_53761244_29-187	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Paer_46164403_16-175	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Paer_15598903_29-190	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Psyr_28868700_31-188	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Psyr_23470466_31-188	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Pput_26988221_31-190	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Pflu_29611996_37-195	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Ecar_50120707_32-188	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Ecar_50119142_33-189	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Ecar_50119143_33-190	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Tden_52006605_43-203	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Xcam_21231332_32-190	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Gsul_39996402_33-188	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Bcep_46319877_18-182	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Rsol_17549582_28-189	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Bmal_53716753_32-194	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Psyr_23469129_31-192	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----
Psyr_28870840_31-193	-----KRLSAA-----SHEFTVMRTDALPGYVLAASLRASANEYSIYLO-----QAT-----F-----VTE-----AE-----VQDOL-----AKIISDALKE-----

Figure B.1 Unedited seed 4HB_MCP alignment with VISSA visualization.

Hhep_32262705_40-203	NRVSS	NASLSEIN	RNALQRYAINFRGSHVDRAIAVR	DVV	IISSEDX	NG	LQQL	SGINALEKE
Rleg_4973017_32-196	TEEVALL	NDKLGAMNDVNS	VQKRFAINYRGSHVDRAIAIR	DVT	LVTSDDE		RKTAE	ALIGGLAAS
Rrub_48763002_15-178	RVNEI	SHSLDVINEV	SVQKRYAINFRGSHVDRAIAIR	DVI	L-VTSA	GB	ADAVV	ATIDKLAGE
Xaxo_21107857_11-168	RVRSI	DQQLTAINEV	SVQKRYAINFRGSHVDRAIAIR	DVV	L-MDDP	AN	RHAAS	QSIDKLAAE
Xcam_21231495_11-168	RVRAI	DQRLTOINDVNS	VQKRYAINFRGSHVDRAIAIR	DVV	L-MNIT	AD	RQAAE	HAIDKLAAE
Wsuc_34557253_30-190	QKVNFI	QDTLRITIDVNS	VQKRYAINFRGSHVDRAIAIR	DVV	L-TQES	TC	LEPTL	BEIKLEDF
Vvul_27361676_33-194	QKVNFI	QMSLAEVTD	NSVQKRYAINFRGSHVDRAIAIR	DVA	M-ARTI	QE	LAREF	BEIRLEKEI
Vfis_59713711_33-194	QKVNFI	QDTLTETMDVNS	VQKRYAINFRGSHVDRAIAIR	DIA	I-ARTI	QE	LSHLE	SEMIKLEEE
Vfis_59714255_1-171	MFLTLIFGIQKVNFI	NSSLTETSIDNS	VQKRYAINFRGSHVDRAIAIR	DIA	I-ARTI	QE	LSHLE	KEINRLQTE
Vfis_59711698_1-171	MLLTILGIQKVNFI	QDTLTETMD	NSVQKRYAINFRGSHVDRAIAIR	DIA	I-ARTI	QE	LSHLE	KEINRLQTE
Rgel_47574745_31-191	LKIVTAI	QCKFADVMDDR	PKIQTAGDRTVNNEVSLAIE	NLF	V-VSEI	AD	VOAQE	SVIANSSAR
Gsul_39996396_31-189	NRMATI	NTDLDMVVKDRWKAET	TFGSISSQINNVARALR	NAL	L-LDDP	AE	VQKEI	ARINEASVS
Gmet_48845935_32-188	UTHLAI	QNDLALISE	PFKTVQAHNEVNOANIVARAVE	NAL	L-LDDP	VO	VQKEI	ARIEAQFAR
Daro_41725108_32-189	RLSTI	NESINDMVSDFK	PKTVLANDIVNNINIVARAI	NAA	L-VKRP	ED	VSKEL	BRVADAKAR
Tden_52008473_33-190	LQNAQL	NAELERVSV	NVNSLASQMRDALDRAVIMH	NIV	V-TTP	WE	KDALE	LRFQRYGEG
Tden_52006559_34-186	LSQI	QRHVDDVAGSRMR	KRLVNGMRDAMQSGAVAVR	NIV	LTLTGA		MAEEA	QRFLAHNAR
Xory_58582465_75-233	RLSSSA	RALIDDIYNQNMVK	IRLNSNDMMNANFRIGTQLR	NIV	LP-TTAE		AEIQSAR	AEIQSAR
Xaxo_21108106_1-132			M-ANSVIATQIR	NVV	LP-TSNE		NLKEI	ENIKNARAR
Xory_58582471_51-209	RGDTSA	RTLVDALYNQNM	KIRLNSNDMMNANYVIAABLR	NVV	LP-TSNE		NLKEI	ASTIQARAR
Xory_58582468_27-183	LAQR	RGMRDSIVKHN	MAILEYIGEMRSASATAINLR	NIV	MP-TTQE		NLGEA	XVIEQRQV
Xaxo_21108103_1-133			M-ASAATAINLR	NIV	LP-TTQE		NLGEA	XVIEQRQV
Xcam_21231323_1-138			M-HQMDTDDTSVIAVQLR	NIV	LP-TSQE		NLGEA	ALIKDRAKR
Xory_58582470_36-191	LAQR	RGTLDTLTNRNM	VIVORLQEMINNAVSVIAVQLR	NIV	LP-TTQE		NLGEA	XVIEQRQV
Xaxo_21108101_1-135			M-EMINNAVSVIAVQLR	NIV	LP-TSQA		NLGEA	ALIKDRAKR
Xcam_21231752_31-192	SNSGRI	NDLSTTLIEREL	LGLSNVKEANINLIYAGRARA	NLL	LASSAE		RQSHV	QNDIKYTAR
Xaxo_21108705_31-192	SNSRI	SALSSSLYEREL	LGLSNVKEANINLIYAGRARA	NLL	LA-SSAE		RQSHV	QNDIKYTAR
Rmet_48769262_34-195	NMGRM	LEWGTIYNSD	DALKAVQDGNILVYASRAQI	ALL	LS-ASTI	GE	RATER	QSEKESLST
Reut_53762135_34-195	NMGRM	ADWGTIYNNI	DALKAVQDGNILVYASRAQI	ALL	SAS-TMGE		RSTEKEQT	LSLAAMDAR
Rrub_48764797_25-186	GLGVTSLAGI	SANRLGLVDG	VQRIQTAELKLLVDVQVIRAR	NMI	LAVSPOE		AAQHE	KITLGLRPO
Rrub_48764795_29-185	LQKLGAI	NGSLVAMVEG	VQRIQTAELKLLVDVQVIRAR	ALL	LA-GPDT		QQQYE	FRIAKEQOQ
Bpse_53721497_29-189	GFGIYESRRV	YTAASVSTVNT	VPSFVVLQACRAFDSMLLIVN	QOV	ESTTADQ		AKALE	PRIAQOARR
Bcep_46310707_33-189	LNHA	SRLSDEIAHVD	PAIHTLDDTSYLLRARVSDLORESL	TE-GGNI	AE	AAKVI	PRIAQYALR	
Bcep_46322311_35-188	E	NHSLLEAMYRD	DSATLLHKTSSERMLVLRERVSQVQI	IS-AGOP	ANABE	ANABE	QHLTHLQO	
Bcep_46315673_18-170		NASLEAMYRD	DSATLLHKTSSERMLVLRERVSQVQI	IS-AGRS		GKEBE	AKRLTLLKO	
Bfun_48786542_29-189		NASLEAMYRD	DSATLLHKTSSERMLVLRERVSQVQI	IS-AGRS		GKEBE	AKRLTLLKO	
Xaxo_21108114_1-134	GLSLRSS	NASLEAMYRD	DSATLLHKTSSERMLVLRERVSQVQI	IS-AGRS		GKEBE	AKRLTLLKO	
Bfun_48787402_62-215			M-ALLGEFERTYEL	AQI	AD-ADDP	KA	VQNYE	BRMDKARAR
Bfun_48788521_31-187	DVWRV	NRANQDTYQNI	LIAAVYIGNAELLITARTLRVLG	SAM	AO-P-DS	AR	AQEQI	CHASEFTRG
Rsol_17548524_27-187	QGMNRG	NEAQHDAY	IVHPSVYALGKSGTAMSRARFGLD	NAM	SN-P-HS	PO	LAQQL	FRAMLLIGE
Bcep_46321947_17-173	QGLVNLRI	NAALKETYSNN	LALGKAEAKLAQARTALR	AL	FE-T-DE	AK	ETDAL	QRTAQIGR
Bfun_48788500_33-185	QGVAIL	NDDVKTLYSER	LASSEALQANVALSRTLRNLV	RIA	ED-P-DS	PD	VQPSQ	RTARELLAR
Bpse_53719442_41-196	QMSRA	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Bcep_46320035_33-188	QMSRA	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Bcep_46312035_33-188	QMSRA	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rsol_17549625_31-187	QMSRH	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Reut_53762140_18-175	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rmet_48769257_19-176	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rsol_17428475_31-189	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Reut_53761466_20-177	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rsol_17548728_31-187	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Bfun_48787377_20-177	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Bcep_46322449_38-195	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Bmal_53716899_85-242	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Sone_24376324_29-192	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Cvio_34104168_23-184	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rmet_48770444_25-187	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Psyr_28870737_37-189	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Psyr_53693339_33-189	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Pflu_48729692_32-188	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Psyr_28872674_33-190	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Psyr_46188178_1-168	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Psyr_28871675_33-190	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rsol_17549061_29-185	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Reut_53760762_34-191	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rmet_48770042_24-188	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rmet_46131652_18-177	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Mmag_46203065_25-186	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
RspH_8250660_26-186	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
RspH_22958341_26-186	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Cvio_34103819_33-190	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
RspH_7532754_24-187	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Pflu_48730667_20-184	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Psyr_46188223_34-196	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Psyr_46187354_20-176	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Psyr_28870740_34-193	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Gmet_48844820_32-197	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Daro_53729524_186-351	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Bcep_46321623_26-185	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Bfun_48782649_1-172	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Reut_53761418_15-174	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Gsul_39995862_33-195	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Gmet_48846722_1-129	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Ecar_50119387_32-189	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rrub_48764497_1-152	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Mmag_46200981_38-196	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Cvio_34102727_35-197	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Sone_24372094_36-192	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Naro_48849030_29-188	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Wsuc_34558217_24-182	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Cace_15893832_34-189	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rcar_50120989_24-182	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Wsuc_345577239_33-189	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rpal_39934745_190-343	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Tden_52007871_1-136	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Sone_50261353_32-187	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Mmag_23013720_28-169	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Rrub_23013426_39-196	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Neur_30249816_32-185	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
Mmag_46201783_20-177	QMSQS	NHALSDTFT	NAMPASVDIGNAEVLAERERLALR	AAA	EMIGTEF		AAPEI	FRARGHRAI
	NRVSS	NASLSEIN	RNALQRYAINFRGSHVDRAIAVR	DVV	IISSEDX	NG	LQQL	SGINALEKE
	TEEVALL	NDKLGAMNDVNS	VQKRFAINYRGSHVDRAIAIR	DVT	LVTSDDE		RKTAE	ALIGGLAAS
	RVNEI	SHSLDVINEV	SVQKRYAINFRGSHVDRAIAIR	DVI	L-VTSA	GB	ADAVV	ATIDKLAGE
	RVRSI	DQQLTAINEV	SVQKRYAINFRGSHVDRAIAIR	DVV	L-MDDP	AN	RHAAS	QSIDKLAAE
	RVRAI	DQRLTOINDVNS	VQKRYAINFRGSHVDRAIAIR	DVV	L-MNIT	AD	RQAAE	HAIDKLAAE
	QKVNFI	QDTLRITIDVNS	VQKRYAINFRGSHVDRAIAIR	DVV	L-TQES	TC	LEPTL	BEIKLEDF
	QKVNFI	QMSLAEVTD	NSVQKRYAINFRGSHVDRAIAIR	DVA	M-ARTI	QE	LAREF	BEIRLEKEI
	QKVNFI	QDTLTETMDVNS	VQKRYAINFRGSHVDRAIAIR	DIA	I-ARTI	QE	LSHLE	SEMIKLEEE
	MFLTLIFGIQKVNFI	NSSLTETSIDNS	VQKRYAINFRGSHVDRAIAIR	DIA	I-ARTI	QE	LSHLE	KEINRLQTE
	MLLTILGIQKVNFI	QDTLTETMD	NSVQKRYAINFRGSHVDRAIAIR	DIA	I-ARTI	QE	LSHLE	KEINRLQTE
	LKIVTAI	QCKFADVMDDR	PKIQTAGDRTVNNEVSLAIE	NLF	V-VSEI	AD	VOAQE	SVIANSSAR
	NRMATI	NTDLDMVVKDRWKAET	TFGSISSQINNVARALR	NAL	L-LDDP	AE	VQKEI	ARINEASVS
	UTHLAI	QNDLALISE	PFKTVQAHNEVNOANIVARAVE	NAL	L-LDDP	VO	VQKEI	ARIEAQFAR
	RLSTI	NESINDMVSDFK	PKTVLANDIVNNINIVARAI	NAA	L-VKRP	ED	VSKEL	BRVADAKAR
	LQNAQL	NAELERVSV	NVNSLASQMRDALDRAVIMH	NIV	V-TTP	WE	KDALE	LRFQRYGEG
	LSQI	QRHVDDVAGSRMR	KRLVNGMRDAMQSGAVAVR	NIV	LTLTGA		MAEEA	QRFLAHNAR
	RLSSSA	RALIDDIYNQNMVK	IRLNSNDMMNANFRIGTQLR	NIV	LP-TTAE		AEIQSAR	AEIQSAR
			M-ANSVIATQIR	NVV	LP-TSNE		NLKEI	ENIKNARAR
	RGDTSA	RTLVDALYNQNM	KIRLNSNDMMNANYVIAABLR	NVV	LP-TSNE		NLKEI	ASTIQARAR
	LAQR	RGMRDSIVKHN	MAILEYIGEMRSASATAINLR	NIV	MP-TTQE		NLGEA	XVIEQRQV
			M-ASAATAINLR	NIV	LP-TTQE		NLGEA	XVIEQRQV
			M-HQMDTDDTSVIAVQLR	NIV	LP-TSQE		NLGEA	ALIKDRAKR
	LAQR	RGTLDTLTNRNM	VIVORLQEMINNAVSVIAVQLR	NIV	LP-TTQE		NLGEA	XVIEQRQV
			M-EMINNAVSVIAVQLR	NIV	LP-TSQA		NLGEA	ALIKDRAKR
	SNSGRI	NDLSTTLIEREL	LGLSNVKEANINLIYAGRARA	NLL	LASSAE		RQSHV	QNDIKYTAR
	SNSRI	SALSSSLYEREL	LGLSNVKEANINLIYAGRARA	NLL	LA-SSAE		RQSHV	QNDIKYTAR
	NMGRM	LEWGTIYNSD	DALKAVQDGNILVYASRAQI	ALL	LS-ASTI	GE	RATER	QSEKESLST
	NMGRM	ADWGTIYNNI	DALKAVQDGNILVYASRAQI	ALL	SAS-TMGE		RSTEKEQT	LSLAAMDAR
	GLGVTSLAGI	SANRLGLVDG	VQRIQTAELKLLVDVQVIRAR	NMI	LAVSPOE		AAQHE	KITLGLRPO
	LQKLGAI	NGSLVAMVEG	VQRIQTAELKLLVDVQVIRAR	ALL	LA-GPDT		QQQYE	FRIAKEQOQ
	GFGIYESRRV	YTAASVSTVNT	VPSFVVLQACRAFDSMLLIVN	QOV	ESTTADQ		AKALE	PRIAQOARR
	LNHA	SRLSDEIAHVD	PAIHTLDDTSYLLRARVSDLORESL	TE-GGNI	AE	AAKVI	PRIAQYALR	
	E	NHSLLEAMYRD	DSATLLHKTSSERMLVLRERVSQVQI	IS-AGOP	ANABE	ANABE		

Cvio_34101575_31-187
 Gsul_39996476_32-189
 Pput_4235480_32-187
 Pput_32469921_33-187
 Naro_48850981_34-191
 Rmet_48770115_21-183
 Xory_58582478_34-191
 Xcam_21231317_34-191
 Dvul_46578600_31-194
 Cvio_34102891_42-206
 Retl_21467290_23-187
 Pflu_48728799_32-190
 Psyr_28870443_33-190
 Psyr_23472455_33-190
 Pflu_48729489_15-172
 Pflu_48733189_32-188
 Pput_26987059_33-188
 Psyr_28867696_32-188
 Pflu_48732089_27-183
 Ctet_28211515_29-189
 Mmag_23011382_54-210
 Mmag_23013949_31-187
 Sone_24373012_30-190
 Paer_15596448_31-190
 Paer_53727587_31-190
 Paer_15596805_32-188
 Dvul_46580384_49-207
 Psyr_28868132_32-190
 Psyr_23472827_32-190
 Gsul_39995789_34-190
 Rub_48764903_34-190
 Bjab_27377659_30-188
 Bjab_27378042_32-190
 Rpal_39937697_28-190
 Rpal_39937696_30-192
 Rpal_39934710_32-186
 Psyr_28871756_32-186
 Psyr_46189040_15-171
 Iloi_56459719_31-188
 Bbac_42522500_32-194
 Bbac_42524709_29-190
 Rmet_48727490_1-153
 Reol_17549248_31-188
 Reol_53761145_32-187
 Linn_16799806_33-190
 Lmon_16802765_33-190
 Lmon_46906974_33-190
 Dhaf_46586096_14-173
 Bsub_16804022_25-184
 Ctet_28210899_35-191
 Cace_15896748_32-189
 Cace_15896020_28-189
 Cace_15896019_33-189
 Cace_15894666_33-188
 Cace_15896638_34-188
 Ctet_28211181_69-221
 Ctet_28211184_33-183
 Dgig_4235392_33-193
 Esp_4614138_32-187
 Bcer_52144881_34-189
 Bthu_49476793_35-189
 Bant_47525640_35-189
 Dhaf_23121705_34-189
 Bagr_29170613_36-187
 Dhaf_53685321_20-174
 Dhaf_53685000_1-147
 Blic_52002118_32-188
 Bcer_52142147_38-193
 Bcer_30021489_34-188
 Cace_15893416_33-189
 Mlot_13488383_33-191
 Cglu_41326927_358-461
 Xcam_21231331_61-200
 Xaxo_21108113_80-225
 Rleg_15072891_30-186
 Rleg_2665910_26-185
 Vfis_59712649_32-186
 Wsuc_34557328_34-190
 Wsuc_34558241_35-189
 Cthe_48860112_13-169
 Cthe_477735_1-125
 Vpar_28900346_32-187
 Vfis_59713352_32-189
 Vfis_59713351_32-191
 Vfis_59713353_32-185
 Ppro_54303603_31-193
 Vpar_28900853_31-191
 Vvul_27358478_31-191
 Vcho_9655802_32-191
 Dhaf_23113288_34-189
 Dhaf_39997674_35-189
 Dvul_46580722_34-190
 Gsul_39996400_29-188
 Gmet_48847252_32-190
 Gsul_39998033_33-189

-----HTEGTA-----LGDFATTYNDRVVCLKQLKIVGSGYAVGIVDNA-----OKL-----RN-QSOT-----VAGFL-----ANLQAKQLB-----
 -----NTARTA-----NNGLDTVYDRVLPLKDLKIIADMYAVNIVDVS-----HKV-----RN-GTIT-----NTEGR-----KSVVEAKKT-----
 -----LQRC-----VASLNTVYLDRLVPLRLDLKTIADLYAVKIVDSS-----HKA-----RS-GMT-----VAAQE-----QEVKDAGRO-----
 -----ERG-----VASLNTVYLDRLVPLRLDLKTIADLYAVKIVDSS-----HKA-----RS-GRMT-----VAAQE-----QEVKDAGRO-----
 -----GVGAN-----KEQVVFIDENITVPSIAQIGDVSIGVMEARFAMS-----KLT-----LA-SRTG-----VA-----ROEQO-----STINEKRAA-----
 -----AFSINRLSAR-----NVAEISGWLISNRIAGDLNNSISDFRATEG-----ELL-----TA-REG-----KD-----RGRAR-----SGLAVDOEE-----
 -----YSGLASI-----NIVTRDLANGTMSVREAGDLRGLGEYRNAAV-----QNL-----VR-ASL-----SV-----KQEAQ-----VRSNKLNGK-----
 -----YSGLASI-----NDVTSGLAGNTVPSVREAGDLRGLGEYRNAAV-----QNL-----VR-ASL-----SV-----KQEAQ-----AVNGKLKRO-----
 -----LAFSR-----DTAAHSGVITVLPSTAIAGMGOGLARMKVRQI-----SVL-----AA-EAG-----EA-----LAEGE-----NLTESSKRA-----
 -----BOYLMEV-----YGANFGNVNAVSLKVLGDLRRAEQEC-----TACG-----RLD-----YD-ADK-----AE-----RLKVE-----NKVLAAREA-----
 -----QSVISVKKLSI-----PSNISEVANSNLPSPDVIKINWITADYRILOQ-----RLV-----TN-SSDS-----TS-----LAHNI-----VTRQRMED-----
 -----QOMSSI-----PDSEYAVEITOWLPSIRGDEIREIMLRITISI-----RMA-----L-QDS-----AN-----LATYV-----QMDTRDRE-----
 -----QMSI-----PDSEYAVEITOWLPSMRYVNDIREIMLRITISI-----RMA-----LD-TDQ-----AS-----LFTYR-----QQLDVLQGE-----
 -----QMSI-----PDSEYAVEITOWLPSMRYVNDIREIMLRITISI-----RMA-----LD-TDQ-----AS-----LFTYR-----QQLDVLQGE-----
 -----LQMSNN-----RAQSDEVDNWLPSVMVAGEMSQDMLRLRALTN-----ALL-----L-NRD-----QA-----LDQNY-----AKINDLRGV-----
 -----NRMSI-----RQASLEMDSTCLPSVTQLAVVTENVLRRLILSH-----KIL-----V-NRD-----AG-----LQEAQ-----TRIGVLYDV-----
 -----QMGNI-----RQAGVAIEQVSVPSIKILDELTAIINLRMTLSI-----KIL-----LN-REP-----AT-----QROTI-----BMDQRNSG-----
 -----QOMNKI-----RGATEDLANGNVPSIKSLDRFAEVSTIRLVLSI-----KIL-----LN-RDQ-----ET-----QOKTI-----DILAMNRQO-----
 -----QOMSKT-----RGAEDITSSSVFSTINDEFTQTLRLRLVLSI-----KIL-----V-NRE-----DV-----QOKTN-----DILDMNRQO-----
 -----KSLAE-----NEKSTIDIAQCGTIGIYSEELNIMTSDRFLFEX-----GHI-----IS-KDC-----GR-----MKRRE-----KSMEEKNKE-----
 -----LADI-----SOENRNLGNILPSVRYAABLGNVDIVRVSVVA-----NHV-----DY-TDE-----DR-----TAASE-----KSLAKARMA-----
 -----NRLAAV-----NQLSTDMENWLPSPVDSARELDGLLAKQRAITV-----RHI-----DT-TDQ-----AQ-----MAEVE-----KILAAQHSI-----
 -----KMQVT-----NEQSTVISSNWLPSRVYIPELNGQTDALRLVLSI-----KIL-----LS-LDN-----DQ-----MRETE-----RELEKIKLA-----
 -----NRMGSI-----NRAAKDIGEVLWLPSEVSAQSLGLMSELRLGEN-----NHV-----LH-HDA-----TR-----MRQOE-----QRMDEVAT-----
 -----NRMGSI-----NRAARDIGEVLWLPSEVSAQSLGLMSELRLGEN-----NHV-----LH-HDA-----TR-----MRQOE-----QRMDEVAT-----
 -----NRMGSI-----NEASSDINWLPSPVSEAEALNVLAELOVQI-----AHV-----LA-ADP-----SS-----KRFLE-----SMTEIGKE-----
 -----NRLGLI-----HDDVQELIASNWPSIKILAKMQGDMNRIRRQOI-----GHI-----TA-TDA-----SA-----LASTE-----SNIKELKQO-----
 -----LIQLGV-----NQAQADIKENMMPSMRAAGSMRFFAANYRLKEN-----RHI-----AA-DAP-----P-----KAQME-----QEAARSKO-----
 -----LIQLGV-----NQAQADIKENMMPSMRAAGSMRFFAANYRLKEN-----RHI-----AA-DAP-----P-----KAQME-----QEAARSKO-----
 -----ELSRV-----NETQDMAENWIPSLNIAISAMQDLFASYRLEI-----GHI-----LE-VEA-----S-----OKTVE-----SRMAGLVKE-----
 -----OLRAV-----SGATEIIVENWLPSPRALAALEQVIEHRRFEM-----SHI-----MT-TDA-----S-----MAABE-----BRIRKQREI-----
 -----LMQNI-----NAHVEVIAEVLPSVRALGSLRADINELRVALL-----LHL-----MQ-DSAE-----S-----KQAAE-----KRIASLOGR-----
 -----KMRAN-----NANTDITISWMPSPVRVLGELRASVITYRSVVR-----BHM-----LQ-ETLE-----S-----KLAME-----KIAVTTSST-----
 -----LVLKLSH-----NNTVETITNOLPTVTVLGEARVLSYRMLKE-----DHL-----LE-TND-----S-----MAAIE-----KQIDASIGK-----
 -----GLKMRIT-----NNTSVDIATNWLPSVRVLGDIKTILAMYRVTL-----AHA-----ME-TTAE-----S-----KATAA-----KRIAGVLEI-----
 -----QMRIT-----NTSVDIGINWLPSPVRVLGEMRTNVALYRNTLE-----LHL-----LE-ATPE-----S-----KLATE-----KQVAAFOEA-----
 -----QMKSI-----ADTEKDVLEINWLPSPIRQTAAMNSITVLKRVLETO-----RAV-----A--DQ-----QI-----KQVFI-----KKEPARKA-----
 -----KMKGEI-----RASMESLEKDAVASIIQANKISSATLRLRLDAR-----ALI-----R--TDP-----QA-----QITIV-----ERLKAAREI-----
 -----NRIET-----DNGITVSEKOLVPLESEHLEQFLLVRIHSA-----NVS-----VS-VND-----QS-----LATYV-----KILADITGO-----
 -----EMREI-----ASSYKVTGDI-----PNIESADQYMMFRCVRISSLSG-----GL-TDQ-----S-----KADTI-----QVDEANIAS-----
 -----YFLKSI-----SABYNIVAHENLPSLKLADLSTIRELRLHVRSGI-----L-----GN-TDE-----P-----VNIIV-----KXSGEQIAN-----
 -----MLR-----DGYVKAIYE-----NTAFRELABETRARATDIRRLIM-----KLV-----LUR-GRPE-----P-----KQAI-----KAIHNLRA-----
 -----VGLMRA-----NGNTDAYIGNTITVIAQLAEVRAQDLRLQMRRTAL-----KDF-----KDF-----P-----VTAAL-----QAITDVEG-----
 -----VTSSE-----NENMHDAEYQNTVIAIGLADVRAQDLHIRRLIM-----KIQ-----T-SEDE-----P-----SSAAN-----ALVRVLEI-----
 -----LGRHY-----ATLSDNMVNNVAPMKIEIAKIQTNMAQINIDIT-----IMP-----D-TING-----KSTLI-----KIDITLYAE-----
 -----LGRHY-----ATLSDNMVNNVAPMKIEIAKIQTNMAQINIDIT-----IMP-----D-TING-----KSTLI-----KIDNLYAE-----
 -----LGRHY-----ATLSDNMVNNVAPMKIEIAKIQTNMAQINIDIT-----IMP-----D-TING-----KSTLI-----KIDITLYAE-----
 -----LNLGV-----NANIKITQDGIQFILLLEDLNLKSFAGAAELI-----GVV-----MKSGVAE-----SAVVGDSR-----RVIAQOET-----
 -----STGVLMQHI-----IQKTDITNKWDIGIKGITSINYVTEHLSSEK-----DPL-----IYTDKSK-----MOTLE-----QEMQINQI-----
 -----NMQVY-----NNSGERLYHE-----LIGINSVRNIKENFLTIGANT-----LMA-----SOENLGR-----VSYLE-----NDVKRLPNI-----
 -----SYEKMTQV-----NDNTIITVSESLMKISITKIDIRAHMADLTSGSI-----LIV-----NPR-----KSVIKDTI-----SDMGLTEK-----
 -----SIVGILDMSKI-----DINLESMEYIDIM-----RTNILEYELKTNVDIKADTN-----LIL-----KS-TNEV-----KSDNIDIKIALKVI-----
 -----LAKNGV-----NKNLDNIYNVDMKGNILQIPIGVIEVRAIT-----L-----MD-PMK-----SK-----LDSIV-----SIBDIDVQK-----
 -----LSLKTV-----FSNLQNIYKVDLKGNDLQLKADMETRADMI-----KIK-----DKN-NDK-----K-----VDDLII-----KEISDLNKI-----
 -----NNMKV-----NNNVKSMYNNLMPTSIDIGIKSEFLKIRLDAT-----NAA-----R-STV-----SSEYA-----SNIKNYDN-----
 -----CATNKV-----NNNAVMBEELLPVNIITGMREGFLSVRLNSA-----NAK-----TDE-----NEKYV-----KSTIEHDI-----
 -----FEKV-----NNNVTIYDKGILPVIIRINMREFTLMRLNCT-----NAI-----LAY-----P-----KMKHI-----NLIESNEE-----
 -----SMQTI-----NALLEIYENNKPIVHLADANAEEIKFGRNOV-----RIF-----IS-TND-----P-----ACEFV-----KRGENDAN-----
 -----LNQI-----SNASOAMYKENLIPVOEVAQIRIDTRALDSFLV-----BMM-----ITKDEAR-----P-----IEELG-----AQIDQROAQ-----
 -----LNQI-----HKQLQTVVYVDRLOPKIWLGSIESSLYQEFYSYV-----ELI-----IT-EDN-----S-----RTTIL-----NKINDTNKE-----
 -----LNQI-----HKQLQTVVYVDRLOPKIWLGSIESSLYQEFYSYV-----ELI-----IT-EDN-----S-----RTTIL-----NKINDTNKE-----
 -----LNQI-----HKQLQTVVYVDRLOPKIWLGSIESSLYQEFYSYV-----ELI-----IT-EDN-----S-----RTTIL-----NKINDTNKE-----
 -----DKIA-----NNNLQMTYED-----LIPVKIAYDTRSDIRAMNGLI-----PSM-----ITADKCK-----MOTLI-----PQIQEQOG-----
 -----KYS-----GERMOAMYEBGLIPVKIYNNRQDIRAIDGLI-----KMI-----LDAP-----AV-----FAQYV-----BELDQRIAS-----
 -----L-----LAAVENNQNNLIPVKIYNNRQDIRAIDGLI-----KMI-----LDAP-----AV-----FAQYV-----BELDQRIAS-----
 -----L-----MYQNNLIPVKIYNNRQDIRAIDGLI-----KMI-----LDAP-----AV-----FAQYV-----BELDQRIAS-----
 -----VGLSKA-----SKGSEMITQDQILIPNQLFARLKANNLDLDTYKI-----LIM-----YT-KDS-----P-----NSTIQ-----ANIKERNEE-----
 -----GLERG-----KSTSSMYEDNLPIEWIGIVGSENFYHVNMMFM-----LIM-----YS-KDS-----P-----NSELT-----REMDGIRKE-----
 -----SLERG-----KSTSSMYEDNLPIEWIGIVGSENFYHVNMMFM-----LIM-----YS-KDS-----P-----NSELT-----REMDGIRKE-----
 -----LNLNKV-----NNNNTETVYKMLVPTGLVHSISNDFTMKARV-----DVA-----TEQMDK-----MOKFS-----SDFNAASS-----
 -----LOR-----SSEATEVEK-----PDYHSEITIMIGOMTAHSLG-----DVA-----QDNPSI-----KSTIE-----SATTNRQO-----
 -----FT-----LNG-----
 -----T-----PAA-----ASVLDNVP-----VGALQSIATDLVOLRNTOR-----AQL-----AA-ADP-----AR-----IDAVD-----ERLGNLRQR-----
 -----PPSFAAP-----AA-----VSVLLDN-----VPVTAQLSIAVLDLAQLR-----DOR-----AQL-----AA-RDA-----AR-----IDAVD-----ERLGNLRQR-----
 -----STISTI-----RANTEQIGTFWMQRLVTAAREIKDNFLDLKTVYA-----QYL-----EDTAE-----P-----REVGO-----OKIDAGTA-----
 -----VLSLSTISTI-----RANTEQIGTFWMQRLVTAAREIKDNFLDLKTVYA-----QYL-----LE-DTAE-----P-----RMIGQ-----OKIESAGAA-----
 -----LNLRSI-----NDNVSTIITNKSLSIASISLLKGIQVDITKVRKDEI-----SLI-----ENAGHS-----INDWL-----KDLQWRAB-----
 -----YSTGKI-----NDADTILYK-----VSIALVSELMTQGLYRVAFYRYKTN-----LIS-----LSDY-----P-----KVGQDFDY-----
 -----ANTI-----NDNNTRLQA-----VPLGVIKGLNGSLEEVGRYFY-----RYV-----GAIIT-----YDKIK-----SVSNWFAKI-----
 -----NINNM-----SQADAELYEKNTLGINYAAGASLRFQRMRYNTA-----KLL-----Y-DAC-----VSKGI-----KRIQHEVEN-----
 -----M-----YNTA-----Y-DAC-----VSKGI-----KRIQHEVEN-----
 -----SEIKMI-----EGKLTVPFSEITVPSVLLVKNTEIELGLRKDEI-----SLI-----LN-VNE-----P-----FMEV-----AGLESKQO-----
 -----OELNKV-----ESEVINFTDSTVPSVLSVEAMVFEMSVLRSSQV-----ASL-----TY-KEY-----SG-----LWALE-----KRONITLAN-----
 -----OELNKV-----EAEVINFTDSTVPSVLSAEELYFEMNAYERNQV-----AAL-----NY-ETD-----LALGI-----LILSKQENI-----
 -----OELNKV-----ESEVINFTDSTVPSVLSVEDVYFEMNAYERNQV-----AAL-----TY-QEK-----LPLLI-----ALISKQBAQ-----
 -----NELNTN-----NENLLNYTDD-----LPASEQVDSIHQOANLRSSQO-----AVE-----LN-EGSK-----P-----IAQKI-----NQNKETIKM-----
 -----SKELKEI-----KSELNLYTDDTLFAMEKRVDAIRDDLSHWRSSQO-----ATY-----LYK-DAC-----P-----IRNKI-----ASNIEREK-----
 -----STELNT-----KSELNLYTDDTLFAMEKRVDAIRDDLSHWRSSQO-----AVE-----LN-NDEN-----P-----IKQTI-----PNEGIRHE-----
 -----SELNKV-----KSELNLYTDDTLFAMENVDIKOMSYWRRTQO-----AVL-----EMKDEQ-----LQPTI-----ERNNRQV-----
 -----NTRL-----QCMQELLYTYOTLPLLELRVHGSFEENRAYIR-----DII-----L-EDF-----P-----LTHLI-----KALEGNRI-----
 -----LRR-----QSDRLRYEKITVPMHDLAEMSVAFQVRKINLE-----NAV-----P-ATD-----P-----QALVI-----PILKLEBV-----
 -----KLHA-----QVADTRLYE-----M-----VPLDAGNAAVAFQRTINME-----DAL-----NAPT-----P-----KAKAR-----PILRLGRT-----
 -----NLGKTIQI-----EKADTEMELNNAKPMGFIITAVAFQIRVNYI-----LIA-----L-EQT-----P-----KIKFS-----NRIKELQI-----
 -----SKVNTI-----ANEADEMTYNTKPLGVGVQVIAFQARVNVN-----SMI-----DSDS-----P-----AQANA-----NAIRKLYK-----
 -----LKRIV-----ETAGTEMIDLVTEPLGTMGGVIAFQARVNVN-----SMI-----LD-DN-----P-----RAQANANSIAKPYK-----

Ecol_2506837_34-190
 Ecar_50120625_27-192

-----LAQAATHYK-----KSKM-----APL-----EM-----VATSRNIDERYKN-----YTALTELH-----YLD-----G-----RGAYFA-----P-----P-----
 -----ANDYFIRYE-----KMF-----Q-----DARQ-----DOST-----SRGVKENYVAI-----NAALTELIG-----ELN-----A-----SG-----FKKFE-----P-----P-----

Figure B.1 continued

AEKNWQYE	ALP-R	DPRC	SEAA	FLEIKRTYDYN	HGALAEALIC	ALG-A	GK	INSEF	Y	Y
AEKNWQYE	SIP-R	DPRC	SEAA	FLEIKRTYDYN	HGALAEALIC	ALG-A	GK	INSEF	Y	Y
AEKNWADYE	ALP-R	DPRC	STAA	AAEIKRNYDYN	HNALAEALIC	ALG-A	GK	STSYU	Y	Y
AEKNWADYE	ALP-R	DPRC	STAA	AAEIKRNYDYN	HNALAEALIC	ALG-A	GK	INSEF	Y	Y
AEKNWADYE	ALP-R	DPRC	STAA	AAEIKRNYDYN	HNALAEALIC	ALG-A	GK	INSEF	Y	Y
AEKNWADYE	ALP-R	DPRC	STAA	AAEIKRNYDYN	HNALAEALIC	ALG-A	K	INSEF	Y	Y
AEKNWADYE	ALP-R	DPRC	STAA	AAEIKRNYDYN	HNALAEALIC	ALG-A	GK	INSEF	Y	Y
ATVNMHWG	QV-	HDE	LEQFEAFQKV	QCFIDFLRL	SVR-B	GL	EL	EISPAAREED	D	N
ATVNMHWG	ATSE	NAVE	NA	LEQFEAFQKV	QCFIDFLRL	SVR-K	ND	ASPAAREED	D	N
EDVENHNT	KLP-	GEQE	NA	LEQFEAFQKV	QCFIDFLRL	SVR-B	GL	EISPAAREED	D	N
EDVENHNT	KLP-L	DOKE	AG	AAEIKRNYDYN	HNALAEALIC	ALG-A	GK	INSEF	Y	Y
EDVENHNT	SIP-F	TEPM	NA	AAEIKRNYDYN	HNALAEALIC	ALG-A	GK	INSEF	Y	Y
FRKQAEITYD	PJP-R	TPEE	GR	YQATKPLFFPY	YASMBKVIA	YIT-SAD	PAK	YAL	ABESQD	G-R
FRKQAEITYD	PJP-R	TPEE	GR	YQATKPLFFPY	YASMBKVIA	YIT-SAD	PAK	YAL	ABESQD	G-R
QAQRVDRTYD	NUT-S	SAPE	NA	VDDYQKQNETV	YATKQVIT	YHN-Q	NO	YAE	APRLSE	G-R
AKTRVAFYE	A-M-D	DSET	GR	VGDFFNRPRDKR	RATTBERRLS	SID-S	NR	ASA	AVQISN	G-R
SDQGFNAYE	SA-M-K	TPAD	NA	DNELNARVATY	INGLOPMLR	YAR-N	GM	FEA	TINHNS	G-R
SDQGYRAYE	NRPVK	TPAD	NA	DTLGNQRFQAY	ITGMOPLMR	YAR-N	GM	FEA	TINHNS	G-R
SDQGYRAYE	NRPVK	TPAD	NA	DTLGNQRFQAY	ITGMOPLMR	YAR-N	GM	FEA	TINHNS	G-R
SDQGYRAYE	NRPVK	TPAD	NA	DTLGNQRFQAY	ITGMOPLMR	YAR-N	GM	FEA	TINHNS	G-R
IQESLKRYA	KIS-I	YPAE	NR	LDDVITASKSY	MDETARLKS	NAD-S	GA	D-E	TLTVND	A-H-L
EDNTHALE	Q	SEAS	QDVYFATRRFLANKNY	VSNKNEPNT	ALQ-C	GR	ANE	GRDFPM	K	E
EDNTHALE	ATPKL	NAEX	NA	YDADRFLANKNDAM	KDANSEFLRNIR	ELGUVFN	PENTA	GN	ANL	ATAHAK
VINTIQLQ	SVT-	B-G	SE	LQNFISLWTAY	KSLAQIUVS	LSL-E	NN	KGR	AFETSI	S-K-G
ERERLSDYQ	RIP-M	SKEC	OP	STELQNLKLSYD	LNDGLVPLTN	AMR-	GD	YAG	ENQITL	R-L
LDANWRELD	ALP-H	SPRE	OC	LNAQAQARRTA	DNAAQELRA	LIL-R	RD	LTA	UGRFAD	T-R-L
AKKQADRYE	AA-SV	TGEE	NA	LKKIEAGLAGY	KNATQOVIS	MQR-	MD	FLA	QSPFWM	G-R
VEKGFSEYIL	ARV-Q	SGTE	ND	LIALQDAYKAF	MGLOQEOIA	LIT-Q	NK	LDE	ARTLAN	N-M-L
VEKGFSEYIL	AV-Q	SGTE	ND	LIALQDAYKAF	MGLOQEOIA	LIT-Q	NK	LDE	ARTLAN	N-M-L
FAKALEAYD	PBL-D	SPGE	NE	LEGVDYTFQGY	ARHABOVHA	LIT-A	GQ	EDA	SRLLAW	N-M-L
FOKAVEAYD	PLI-T	EDDE	NA	VEGLKSTYQGY	IEBAEKVYIV	LIT-E	NO	AGA	SRLLAW	N-M-L
FOKAVEAYD	PVI-T	EDDE	NA	VEGLKSTYQGY	IEBAEKVYIV	LIT-E	NO	AGA	SRLLAW	N-M-L
MAGYVOAYE	AW-D	SESE	NR	LNAVEAAWHEV	VRANHERFT	PAV-R	LV	SDG	SVQPEY	S-R-M
TRVQKQELE	A-L-S	HAES	NA	FADISGRKKAY	LDQBAALR	ALA-A	GD	VTA	YELLER	S-R-M
TRVQKQELE	SIL-T	DAER	NA	FADISGRKKAY	LDQBAALR	ALA-A	GD	VTA	YELLER	S-R-M
TRVQKQELE	SIL-S	TPEE	NA	FADISGRKKAY	LDQBAALR	ALA-A	GD	VTA	YELLER	S-R-M
TRVQKQELE	PML-T	SEAE	NA	FADISGRKKAY	LDQBAALR	ALA-A	GD	VTA	YELLER	S-R-M
TRVQKQELE	PML-T	SEAE	NA	FADISGRKKAY	LDQBAALR	ALA-A	GD	VTA	YELLER	S-R-M
TRVQKQELE	QOI-	DTAE	TS	LKQODTREFY	NAADKLAM	WV-S	DEQ	REL	UNALIT	T-K-L
TRVQKQELE	K-T-V	SEET	NA	VATLQVRPAP	NNMKKAIT	LAM-T	NO	HEA	TRPML	T-K-L
TRVQKQELE	D-T	AEKH	NO	VATLQVRPAP	NNMKKAIT	LAM-T	NO	HEA	TRPML	T-K-L
TRVQKQELE	R-V-T	LDGE	NA	LAAMKARVAY	VTAFSEVDR	ALF-A	GQ	REA	QOAGL	T-K-L
TRVQKQELE	VAL	SGMACVSDEE	NR	FATLDGVESRY	GPVALDIVC	LAL-D	GR	RE	ATVMA	T-K-L
TRVQKQELE	GAVAL	SGSATARD	NR	LAKIESVQRI	GPVALDIVC	MAI-L	GR	NQ	ATVMA	T-K-L
TRVQKQELE	TRL	AGGVSKRA	NR	GAETRIENAT	GPVALDIVC	JAA-E	GQ	REA	QOAGL	T-K-L
TRVQKQELE	MA	ADGVSKRA	NR	VAKIASVETKI	GP					

KQKADALV	AMP-S	EQG	Q5	RAEVDLSIPKV	KENNNKVY	AVI-N	GH	SQC	ALPLVL	Q-A
KQNRKRLI	ARP-P	SNAA	QAO	RTQDIAREKA	RENQOVM	GLI-N	DK	PDE	ALKVLN	Q-A
KQNRKRLV	AP-P	SNAA	QAM	RQQDIAREKA	RENQOVM	GLI-N	YK	PDE	ALKVLN	Q-A
KEQTRQTLV	AS-P	SEEA	Q4	RDKIDAA5AA	ALNAQVAB	GLI-N	SK	PDE	ALMLN	Q-A
KKDAHKLJ	TE-A	TVFV	Q4	RAKIDASNKA	ALNQOVIU	LDI-S	GR	TNE	IMPLI	Q-A
KKDAREKLV	TIS-A	TPAA	Q4	RERIDANNKA	ALNQOVIU	GLI-A	GK	ADQ	ILPLIL	Q-A
KVDMYTRAR	AST-E	DEEG	Q1	LAQDRANQKY	LDERGREI	KAN-N	EA	LRE	INPELA	LESAV
KVDHMAKAR	ESF-Q	SEEG	Q4	LAQDRANQKY	LDERGREI	KAN-N	EA	LRE	INPELA	LESAV
LETRVSKAS	DSF-T	TEEG	Q4	FKQHEITLPPF	BERMGKVE	ITG-K	QF	LDTS	FESIVS	Q-A
KQVADKFE		QPEG	Q4	VKQQTALAPF	BERMGKVE	ITG-K	QF	LDTS	FESIVS	Q-A
ITNLIQAV	SAA	TEEG	Q4	WEAFVWANGF	ALFDBKVRG	SVI-A	GE	FLA	AQBLSV	Q-A
ITRRDQJLA	ALA	TEES	Q4	ISAFITPLQOQ	ATLOSQIUA	LAR-S	GR	KOE	ATALSR	Q-A
VDAQFARYE	TUL-S	NDKJ	Q4	LAADRANVSOI	DAVRNVEIA	LSR-S	GR	KOE	AGBLNG	Q-A
SNQNWQAFQ	STPKL	SVCE	Q4	VOELITRYTTI	VKRGVBPFFA	NAR-A	GD	MAF	THAVAF	Q-A
SNDELDAYT	RJHAR	DADE	Q4	FDITQSRRRIT	LQGVFLKMS	OLD-H	DN	ASD	FLBTHR	Q-A
SNDELDAYA	GJHAR	NADE	Q4	FDTLQNRRAID	LQGVFKYAL	OLD-G	DD	GFG	FLDQTR	Q-A
SNDRILAHFV	SRAGT	DADE	Q4	KEMQDRDAPFI	HEAVBPALA	ALR-S	ND	RAA	FQOLQA	Q-A
VATAQDAYA	KIT-T	AGRE	TE	YDEVKILSDY	YKANAALSA	AVR-A	GD	DVT	ANRSDV	Q-A
SDEAWRQFV	DEP-H	BIGE	VF	TEAAGHRDIT	ADAMRAPIT	ALR-S	GD	RDA	AQKIGM	Q-A
SDKWNNAYF	DUP-K	TEPL	Q4	TDPLDAKRTAV	RDDGDKILU	ALR-S	GD	ASW	MDQSLA	Q-A
SEKAWABYT	NUPFS	TPFE	Q4	AGEAKQRDIL	TKIDNTAAVE	ALR-R	GD	KAD	ARGLAV	Q-A
SKKAWDTYR	GJRLS	GPPE	Q4	ADANTKFNAL	HEGDFPMT	ALR-S	HD	PAN	ITPNVR	Q-A
SDMNNKQYM	DIP-R	GPPE	Q4	ADQTVSRREAL	HQGLDAPFA	ITIA-A	ND		QAKVLGAKR	Q-A
SDWNNKKYL	DUP-R	DAQE	Q4	AQDLASKRQIL	QRELDAAFA	ITIN-N	ND	RDR	ILESAR	Q-A
SDWNNKKYL	DUP-R	GPPE	Q4	AQDVAAKROIL	QRELDAPFA	LVIA-A	GD	RDR	LEGEGAR	Q-A
SDTWNNKYL	ALP-R	GPPE	Q4	AQDVAAKROIL	QHECDAPFA	VVG-A	ND	ADQ	LEGEGAR	Q-A
SDAWNNKRYR	ALP-S	TPDE	Q4	ADEADALRTV	FLKRDGAQALM	ALQ-A	GD	AER	TSQTVN	Q-A
SDAWNNKRYA	SUP-Q	NDQE	Q4	ADELARKRADF	MTHGIEPLEK	ASL-A	GN	QEE	AVRLAS	Q-A
SDKGNKTYM	SUP-Q	NDDE	Q4	AGELAAKRAPFE	TGVKPLKE	AMV-A	GN	HDE	ALRLAR	Q-A
SDAWNNKYL	AP-P	HCDE	Q4	SKVEGVTKREA	AGSLRDIUK	ALR-A	SD	RAA	ADATMC	Q-A
SEKAWKAYIL	AVP-R	SAEE	Q4	TQDVSAKREALS	QGVAPMVA	ALR-A	GD	REE	MTSVM	Q-A
SEDWNNKQYK	GUP-M	SADE	Q4	ADRVDAKRTI	FLRQDIEPLVU	ALR-A	RD	AAI	ADKAVN	Q-A
SEDWNNKRYA	ALP-H	DEEE	Q4	ADRVNDAARTAL	LQGVQPSIE	ALR-K	GE	HEE	ADAVYN	Q-A
SEKRWROYE	ALP-R	CODE	Q4	ASRLDAARQAL	LQALKPMIL	AMR-G	GR	RDD	ADRLIN	Q-A
FEGAQAYYA	KINPV	SEQD	Q4	AAQANAARRAL	DEALKPMIL	ALR-A	GR	HDE	ADRLIN	Q-A
FDAALOKYE	KINPV	NDAL	Q4	LOTDEEDMKIL	FLRQDKYLG	ALQ-A	GD	MAS	ANIMIK	Q-A
FEDIMIDKYE	RODTS	DETD	Q4	LAADRAAVKHY	BDQIPAFFE	LLD-T	GD	VAS	AKMIL	Q-A
LAGFSDALG	KUL-V	TPCE	Q4	FDELINKAISDY	QATQNHLY	SVI-A	GN	YEE	AVTISN	Q-A
LAGFSDPDLG	KUL-V	TPCE	Q4	FDELISQAQGY	QVADQRYLL	ALQ-A	GN	HDE	AVATSN	Q-A
ALKRPLDTLK	ALP-A	NDQE	Q4	LEGLSADTAKY	LSLDQJLIL	QIB-A	QD	NDQ	AVARTN	Q-A
LQSRSDPYQ	KJLIS	COGE	Q4	FDDARINKMSYN	ISGLQKVIA	MDV-N	AD	HEQ	AVSFAN	Q-A
LRANSDEPYR	KJLIS	CEAD	Q4	FEENANNKMGY	LDGLQKVIA	LDS-A	AD	HEE	AVSLAN	Q-A
LRNSSEPYR	KJLIS	CATD	Q4	FEEDANSKMGA	LDGLQKVIA	MDG-S	SD	HEE	ALSFAN	Q-A
SVKLRDQLG	KRL-R	TOSG	Q4	FRRLQDTQRTY	CALRARVIA	ATN-A	GD	KEA	ARDLIL	Q-A
NEEKIRGMA	NIP-Y	TKMG	Q4	YRDLDAQATRI	NSRVAIVIA	SVI-A</				

Pflu 48732089 27-183
Ctet 28211515 29-189
Mmag 23011382 54-210
Mmag 23013949 31-187
Sone 24373012 30-190
Paer 15596448 31-190
Paer 53727587 31-190
Paer 15596805 32-188
Dvul 46580384 49-207
Psyr 28868132 32-190
Psyr 23472827 32-190
Gsul 39995789 34-190
Rrub 48764903 34-190
Bjap 27377659 30-188
Bjap 27378042 32-190
Rpal 39937697 28-190
Rpal 39937696 30-192
Rpal 39934710 32-186
Psyr 28871756 32-186
Psyr 46189040 15-171
Iloi 56459719 31-188
Bbac 42522500 32-194
Bbac 42524709 29-190
Rmet 48772490 1-153
Rsol 17549248 31-188
Reut 53761145 32-187
Linn 16799806 33-190
Lmon 16802765 33-190
Lmon 46906974 33-190
Dhaf 53686096 14-173
Bsub 16080422 25-184
Ctet 28210899 35-191
Cace 15896748 32-189
Cace 15896020 28-189
Cace 15896019 33-189
Cace 15894666 33-188
Cace 15896638 34-188
Ctet 28211181 69-221
Ctet 28211184 33-193
Dgig 4235392 33-193
Esp 46114138 32-187
Bcer 52144881 34-189
Bthu 49476793 35-189
Bant 47525640 35-189
Dhaf 23121705 34-189
Bagr 29170613 36-187
Dhaf 53685321 20-174
Dhaf 53685000 1-147
Blic 52002118 32-188
Bcer 52142147 38-193
Bcer 30021489 34-188
Cace 15893416 33-189
Mlot 13488383 33-191
Cglu 41326927 358-461
Xcam 21231331 61-200
Xaxo 21108113 80-225
Rleg 15072891 30-186
Rleg 2665910 26-185
Vfis 59712649 32-186
Wsuc 34557328 34-190
Wsuc 34558241 35-189
Cthe 48860112 13-169
Cthe 477735 1-125
Vpar 28900346 32-187
Vfis 59713352 32-189
Vfis 59713351 32-191
Vfis 59713353 32-185
Ppro 54303603 31-193
Vpar 28900855 31-191
Vvul 27358478 31-191
Vcho 9655802 32-191
Dhaf 23113288 34-189
Gsul 39997674 35-189
Dvul 46580722 34-190
Gsul 39996400 29-188
Gmet 48847252 32-190
Gsul 39998033 33-189

LRKAQTDYE PLI-G SPEE RA X YDQIVQLLQV RQIEDRMK LSR-N-NQ VDE LRNLN I I
IQKYLQFYK KSI-N TRED EE X FNIIVQWQGEY LKLSKEIIT LSR-Q-LK TEE AMNIMR S S
VERTMATYA PLI-S LFGE RA X VETFEREWATJ GHAVTAVIE HSR-E-GR KSI AQEALT A N V
FVKVRDRYA KLI-S SPEE KA-N FDMVVRVMDY DAMTARTID LSR-K-NE NDR AKEVVL G A
LDDSARSYQ QLI-S SNEB QI X FNNFNKLYKEY LSTQCKIIL LSR-E-NK NED AQKILL I S
LARVEREYR PLL-V LDEE RA X LDOFVQRQOEY LEGHAALIA LSR-D-NR TDE ASVLMG G A Q
LARVEREYR PLL-V LDEE RA X LDOFVQRQOEY LEGHAALIA LSR-D-NR TDE ASVLMG G A Q
LTDREGAYO RLI-S GADG RS X MERYLATRGTI LOSQOTILV MSR-S-KP LEB TROYTE G A Q
LQDSARIYE KLI-V DFESE RG-N FEFKPLDHTRQV VRAMDKGME LSR-V-NK NVS AALRLR G S
FETRLATYD KLI-I SDQD RQ-X FSAVSASMSAN LKYSTNLFE LSR-Q-GQ ETR AALRLR G S
FETRLATYD KLI-S SDQD RQ-X FNSVSASMAAY LKYSTNLFE LSR-Q-NL EAQ AALRLR G S
LAHQKQYK PLL-S TPEE KO-X LOEFSTKQGEY LNEGKPVLE LSR-Q-NK AQB AALRLR G S
IAKTQATYE KLI-N SAEB RA X YDRFATQWITJ LALTEQVIE LSR-T-NK NTE ARMDQL G S
IDQTKRYVE P I-T TAEF RS-X FQWTKARLEY LMGVQDVMA LSR-K-V- RFE AHELIQ I I
LAKARQSYE PMI-T SPEE RA X YFWSKLWDDY KKSADVEFA VSR-K-EV-GK FHE AHELIQ I I
IKDSQRRIE ALI-S SPEE RQ-X YRWVEAWGKY RALVPKLLI LSR-R-K YNGQTSVE GASLLS N I I
ENKQLSSYQ KMI-T TPEE KT-X LDEFSEKELNVI LOGAKKAVE MSR-K-SM GOSTAR LGEYLT I I
YQKSLRTYA SMV-A TPEE KR-X LDESLVASWINI TSLAPTGLR MSR-K-BA GELPVA ANDYMA N I I
MKDAVNYTE TLI-A SDQE RQ-X FLAVKSYDDY ASQDLLEP LLK-A-GD TAS EVKLVA I I I
MLKQSAAYA PLA-S AADE AS-X YQVITASQKRE ASLLDITVIE LIO-K-GA NAB AVTFPD N I I
LKESQROYK SMV-C DPOF QK-X SEVSEBELDREY MRLNEGQFNK LVD-A-GL VDA IASIRE N I I
YEVHKQYVL NVN-F KK-E RE-X YEKVDAAMLAH KDLGGKVVO YSR-T-QK PEDRAMLEBIFLYN C
LERHIEAYE IDI-F SAKI RA-S YEFETIAGHWE KSFGEELIS MSANY-GB NESKVSILIR C
LDNAWARYC ADI-AP NOHE DVX ATEFCHRLTVE SGIVRSSLE RIS-S-GD HDD FROWLE G N V
TRNVWKQYV PD IT SDKE RE-V ATGHNIDTLPE NDAANKAVU ALR-A-GN LDR AGQITN G N
MKKGWQAYY PTRVT SPEE RD-X ADQDLTALARE ADLVAQELH LLE-A-GD RAR AARLQE S I I
NDQAIHNEFK KAN-L TAED KKQLAYFEELKADMKSAASSVISDTSS ALD-D-AE LLL AQNRYY G A Q
NDQAIHNEFK KAN-L TAED KKQLAYFEELKADMKSAASSVISDTSS ALD-D-AE LLL AQNRYY G A Q
NDQAIHNEFK KAN-L TAED KKQLAYFEELKADMKSAASSVISDTSS ALD-D-AE LLL AQNRYY G A Q
GDLLLQOYQ MLI-N TDEE KE-X LGEFKEEAVI RDLREKALT AVA-K-ND YALAGDFNO G A
INQKLDNYE KTI-S TDEE QK-X FEOLOTQWNTY MDIHAQITIE SGR-T-ND MDR ARGLV G I I
NNELFNIVE KTI-I TREN KG-X YNRLRKSSTSD RSHINEINN AVK-K-ND YKH AKPHVN G I I
Cace 15896748 32-189 SLD-N DEKD KG-X TFOYLENLNGV NNAKDKFFE TAR-T-GD YNALNQFST I I
DDRLIYNVYK KVI-V MEKO KE-X FEFSEFVNLKK RASRNKVIS YVV-A-GE YEA AKEPED S I I
DNELIKEYK TTI-F TDED RT-X FYQPLNLAKK ROSREKVIA YAK-A-GD YMS ASVEFK N I I
DNKIINEYK NTI-V TAED RK-X FNOVNIELKK GIARSEFVU AVN-G-GD YDH ANSEYF I I
Cace 15896638 34-188 GSG-L DENB NAS ISSFNSNYEEY LNLWTELNS ALKIG-GB TASTKINRA I I
LMKNYKEYT SH-H SDET LKXY LEQFLQCYKNY RDLSEKLEK HLE-R-GB KLTREQ NDS I I
LMKNYKEYT TS-H CDH QOAKY HAFTEFANDY LYMTKEYLE ALK-K-GB KLTREQ NDS I I
ENKALADYT K I-T L TKE BEL YDTFEPANDY LAENAKLRQ MMRN-ED EDV TEPVLE N I I
IRSSLNKEE RGS-T VPT QOSERQ VBNLKSOLVIA DNGLSTIQI LVA-GD KEG SYRYS I I
FEKLQAQYE HTY-T TGESE KK-X LSQYKEKLENI KKHRAQMLD FIR-E-NK LNC AYSPFL I I
FEKLQAQYE NTY-T TGESE KK-X LSQYKEKLENI KKHRAQMLD FIR-E-NK LNC AYSPFL I I
FEKLQAQYE NTY-T TGESE KK-X LSQYKEKLENI KKHRAQMLD FIR-E-NK LNC AYSPFL I I
VEDKLSQIE NLS-L DPOF QE-X FLKTKASYQY PQVKNQIVA FAK-E-AN LNDKAYNLVE I I
ADQTFRYLIQ NSR-L DKEE QE-X FLKTKASYQY PQVKNQIVA LIR-E-NK TDC AYAYYAT I I
VDTLLADYS KAD-M DAHE QE-X FAKIMDTLQI RTERDKAIT MAT-A-GN QDE AFYFYS I I
VDTLLADYS KTIMD TYEC ER-X LPLMDDELQIY RTERSKAVU LAT-A-GK QDE AYTYFA N I I
NNTLMKID Q KLM DNVS EK-X YEFSEKSEYKLI QDISSCOMIS LAV-K-NE NDR AYDYVL I I
NDHLLKQPE AKV-I STKE KE-X YNTFHTFNEI RTQMKRQAE LGK-S-NE NEE AYAYYL I I
NDYLLKQPE AKV-I STKE KE-X YNTFHTFNEI RTQMKRQAE LGK-S-NE NEE AYAYYL I I
PDKVLESFS QTI-L TADG KT-X LSEIKDARKEY VQYAOQAME LSK-Q-NK NVE AMAYVR N I I
VQSRMDDYE ATIVS DIDA QN XAVVKELANQ RLAEATAVUD TSRAALS-GE TPNESPAIKN I I
IKRALDEYE ITAQS QTPE QOOL ITAFRANALAM TADHDEFTV LLA-S-GD YNGAVNVLNKDEE G
IARH G VQO ATVPO GEQA QO FARVQOQLQOY LAQHRQANR ALH-D-GD LHA AQALS G I I
IARHAG FQO ATAPO GEQA RK FALVQQLLGKY LAQHRQANR ALH-E-GD LOY AQALS G I I
LEKVVAEYE KGV-R TERG RE-X INLLKPELAKY RALSEOMIA FEN-D-GK TFE AIRLFK I I
VEKVVTEYE RG-R TERG RE-X INOMKPELAKY RALAEOMIA LEN-D-GK TFE AIRLFK I I
VOTGISSYE SIA-L SSEE RA-S FOEIKETWNOY IKETHSYNN LLS-Q-GN STQ ANEVVL I I
VEKFNKESY KII-V LSKD KE-X HDELIANFKEY KKITLEVLK LLE-Q-GR TEB ADMKTS I I
YEFKPKAYA KAI-S TPEE KAI FEEIVREFTY KSLSLKRE LLE-A-GK TEB AKTFAS I I
FEKYLSLYE STV-I NETD RI-X LQELKALMEKY KSLVDKWEY LVK-S-GK TEB AQKILL I I
FEKYLSLYE STV-I NETD RI-X LQELKALMEKY KSLVDKWEY LVK-S-GK TEB AQKILL I I
LDKYLDQYE K-L-W QORD RD-X PNKVKSAWVY SAFNEYAK LLI-N-N TDSANKTLN G I I
ENMLASYD ATV-A GRES RR-X PDVLAASHKAY SNLFRDFPO LIX-D-RE LDR AHALEN I I
IKALAEAYE TV-I TEHE RR-X FERVANSWQKY LTQLKEFNG LVS-N-GB LKQ AQKELV N I I
IQSDLDAYE ATIV-A SADE RR-X FERVANSWQKY LTQLKEFNG LVA-NK KALDAQELV I I
TEALERAYG ATIV-W FSEE FES IKRMLSGWQY INNI DNFS AMI-L-NN KRA ARSILL I I
ESKELEAYG S IWP GEEQ QI FQRLMRQWQY LVTMDOYNE SM-A-GN KTE ALAVLS N I I
FETELAYG KSVWP GEEQ QI INRMLSLMSGI LSTMDKFNH ALI-A-GG KDA AYPILA I I
INDSLVAYG KSVWP GEEQ QI FKRMLSGNNAI TAVTDQNG TLE-T-QG ADD AYPILA N I I
IQDALATPE QSL-R TPEE NCG FYTLDNVLNE DYHLDQVNE LCR-H-GN KTE AYTILA G I I
ITEHQDNPE KTI-L TDEE RT-X FNEYKARKVY GGYIDNIMO LNS-A-GR VTE AKALLH G I I
LOKDLATPG KALDS ESEB RD-X FDTFETRTQY QALMDKVEY LOE-S-GO HDE AVSIVN S I I
EDDNLPEIE KSI-K SEET KKA YADITKAEIAKE APHLDKIVA LAM-D-GR NDC AVAYMR S I I
IDKQLTEVD ATL-E TEEG KK-Q FATRLTLIKEY GPVREIITA KAT-A-GD RET ALDVNR S I I
IDEGLADPG KSI-L LSKE TROE FADRLNTIKEY APVREIIVA ATI-D-GD RET ALALNR SOG I I

Ecol 2506837 34-190
Ecar 50120625 27-192
Styp 16423100 27-192
Sent 56416317 27-192
Ecol 43218 27-192
Ecol 16132176 27-192
Ecol 26251236 27-192
Sfile 24115585 27-192
Ecol 13364793 27-192
Bjap 27376721 106-259
Cace 15896003 35-190
Msp 48832885 28-180
Gmet 48846841 24-186
Bbac 42522983 1-181
Bbac 18073058 46-208
Bbac 42522982 46-208
Iloi 56461443 33-194
Mdeg 48864484 35-195
Sent 62180191 48-202
Sfile 24113141 45-198

Q---GMQNA MGEAFAYQALSSEKLYRDIVTDNADDYEA
Q---RFQDN FEKAYYVYKAENDRLYQAGIAKNDAYDSA
Q---SYQDA FEKQYVAYMEQNDRLYDIAVEDNNGSSYNQA
Q---SYQDA FEKQYVAYMEQNDRLYDIAVEDNNGSSYNQA
Q---DIRNG FEKQYVAYMEQNDRLYDIAVEDNNGSSYNQA
Q---GYQDG FEKQYVAYMEQNDRLYDIAVEDNNGSSYNQA
Q---GYQDG FEKQYVAYMEQNDRLYDIAVEDNNGSSYNQA
Q---GYQDG FEKQYVAYMEQNDRLYDIAVEDNNGSSYNQA
Q---GYQDG FEKQYVAYMEQNDRLYDIAVEDNNGSSYNQA
Q---GYQDG FEKQYVAYMEQNDRLYDIAVEDNNGSSYNQA
Q---ANRIR VRSKINADLEALQSYDKRAREADQADN Y
Q---QNRQA IFEETKRIVDNMFQSRSENASNTAANAVK
Q---PRFQF IKSLANALLEMNQSNMSEANDRARAKA LTA
Q---FVKEA KQAEIKLSTEGNRAAYEMSOKTKEVDALGQES
Q---SIGTA MHKMSSESMKYINQAAD DKAADEAVTRINN
Q---SIGTA VRWNGAVTIYAKMARDGVIEANSTEAYVKY
Q---ETGKS VRWNSAVTDIYAKMARDGVIEANSTEAYVKY
Q---QOFBA MRNVLEDEQRKEKALSSQSDIIDEVISQSNWTE
Q---TQFDE MRVDITDTITELTRKSVLSDEIAASAAAQIKMM
Q---QLDAA YNHVLLKAIELRTERARLLSEQAYQRT
Q---PLDNN YTDILNKAVKIRSTANQALAEHQRT

Figure B.1 continued

Ecol 16129380 45-198
Ecol 12515286 45-198
Gmet 48847023 33-190
Zmob 56542672 36-189
Npun 53688984 25-201
Lint 45657908 29-179
Pres 27228663 28-198
Dhaf 23120317 31-189
Mmag 23013876 33-194
Psysr 28868215 43-205
Psysr 46188218 30-188
Psysr 28870455 33-190
Psysr 46187691 15-170
Psysr 28870175 31-187
Caur 53795565 28-187
Rgel 47574589 32-189
Tden 52007873 23-184
Ecar 50122563 29-187
Reut 53761194 30-189
Rmet 48772113 31-189
Ecar 50123040 34-191
Ecar 50123255 33-185
Ecar 50123254 32-191
Rgel 47571710 55-210
Rgel 47573506 34-195
Rgel 47571655 31-195
Rsol 17428912 49-203
Reut 46131863 29-194
Rgel 47572127 31-190
Bcep 46324344 24-189
Reut 53761951 33-194
Bcep 46311409 31-190
Bcep 46319966 34-197
Bmal 53723395 36-194
Bfun 48780518 29-192
Rmet 48770099 38-194
Cvio 34102638 19-177
Cvio 34105172 32-189
Cvio 34101405 32-187
Raqu 15077504 33-190
Rsp4 46192461 24-187
Daro 53729525 21-178
Bpse 53722895 48-204
Bcep 46316835 48-204
Bcep 46323416 48-204
Rmet 48771949 47-208
Reut 53761244 29-187
Paer 46164403 16-175
Paer 15598903 29-190
Psysr 28868700 31-188
Psysr 23470466 31-188
Pput 26988221 31-190
Pflu 29611996 37-195
Pflu 48732671 31-188
Ecar 50120707 32-188
Ecar 50119142 33-189
Ecar 50119143 33-190
Tden 52006605 43-203
Xcam 21231332 32-190
Gsul 39996402 33-188
Bcep 46319877 18-182
Rsol 17549582 28-189
Bmal 53716753 32-194
Psysr 23469129 31-192
Psysr 28870840 31-193
Hhep 32262705 40-203
Rleg 4973017 32-196
Rrub 48763002 15-178
Xaxo 21107857 11-168
Xcam 21231495 11-168
Wsuc 34557253 30-190
Vvul 27361676 33-194
Vfis 59713711 33-194
Vfis 59714255 1-171
Vfis 59711698 1-171
Rgel 47574745 31-191
Gsul 39996396 31-189
Gmet 48845935 32-188
Daro 41725108 32-189
Tden 52008473 33-190
Tden 52006559 34-186
Xory 58582465 75-233
Xaxo 21108106 1-132
Xory 58582471 51-209
Xory 58582468 27-183
Xaxo 21108103 1-133
Xcam 21231323 1-138
Xory 58582470 36-191
Xaxo 21108101 1-135
Xcam 21231752 31-192
Xaxo 21108705 31-192
Rmet 48769262 34-195
Reut 53762135 34-195
Rrub 48764797 25-186
Rrub 48764795 29-185
Bpse 53721497 29-189
Bcep 46310707 33-189
Bcep 46322311 35-188
Bcep 46315673 18-170
Bfun 48786542 29-189

PLDNA YTDILNKAVKIRSTRANQLAELAHQRT
PLDNA YTDILNKAVKIRSTRANQLAELAHQRT
FGD VENRAKLLTQYSL GDEHL TAGYFNIR
PAFIC VTSALGTLQNVKQANANKVADMERSYN
CANRQFFTA ATVAFLEVLKMNEDLAAETEHVSAKDV
IRDS IIKTLSYLLKKSEENMO KEENERKY
PAFMT TYASLRTLIESEASAELSFQNDKQF
PAIDP LTLRMQQLSDLELQADAVVRADIVRSERV
DSYQA LMTETQGMVGLKRMVDAARQASQDAATT
LOGDL MDMQVQLLRLINKQSAASAVDAAGASYAQA
LOGDL MDMQVQLLRLINTQSAASAAVEAAGASYEQ
GIANS LMAQLEGLKOLINDASQSEASTGASHTYATAN
VNAEG METALGKLEKINDSEAESSAAATSVYEN
TIAEG LEASLGKLEKINDSEADSSAAASEVYD
VYAT LDREMNLVQ CQARASLDVASSYATARS
PASKV YLAATEDFSNFERELAVQSAAAVKADVQSR
PRLDA YLESGLDLARYQKQADATAGNIHROYESSR
PASNG YLASVEALRDHQASIDQMGKRNINAGASRGD
PAGDA YLAETQKLLDIQRTSIDATAAEINTIYVNA
PAGDA YLVBQKLLDIQRTSIDATAAEINRIYBSARNG
PAQNA LFKVLDAMMASQCCQDNEIVSHQAHQAGSQS
AAQAN VFTALDKMVERQKDLTVELANQSEKEADN
AKQDA VFNALNDMVMNQKLTVEIANQSLKNATNAGS
PALDA LQQRVLDMSQFQARLARETGAEVAARIHQ
PLLSA LVKSAQNFIDLSVDVAKRIVARADAAVANRN
PLLAB LLKAASSYIQYSAERSDAAAKAAEAAVAQDR
PLLDL LVKAVKVYAGYGAELSQRTLQEAEVRA
PLLAA LIHAADHYRDTFSKHSEDLVTAQAAADYAMOR
PLLAA LVKASDDYRDTLSRAVQLTETAEDYLLHRN
PLLSL LKASNDYAAAFANERAAEMVRQSEDOFSNO
PLLAS LLRAVDAYDOLTHRRQRDMEOALADYVSRN
PLLAG LVRAVDAYATYTHREELIAQOQFADRYAMERN
PLLVE LVAATDAFATYSRERAAQLVTDGSKNRYTSQR
PLLAG LVKATNAYSEYTRGRACEMVRRESADHYASQR
PLLAA LISATNAYTDYAHGRQEQLINEFAMHYENORN
PLLAE LIKATNTYASFTRKQDEMIKKLQDDYVAORN
PTNNA FISALLSLRDRQSRLLDKSMQEAAMDSSQA
ASNNT FMAALKDLAKYQEEMNKAVODSQGTYSQARS
PAIDP TSQKFNELKAYQSKVKEAVEASNAAYSQA
TSQNA YLDLSAFANSQDQQLQAEKKAIADGN
TSRAE RTKLLSDLDQRRARNIAEAAAKADAFSSKAQRD
PMQOE WFDVAGDFVLQQRDNDRDVAEINAVKSSVQRT
PAMEE VIRHANVLQGENRRFADQSATLRESVHGTE
PAWTE VVRNANVLQGENRKFPAEQSAKLRESVQDTE
PAWAE VVRSANVLQGENRKFADQSAKLRESGQDTE
DLWST GRKLLQDIDVDNGKVNERASTGIVDSVVTAKVSI
CHWAE VVRLAQTLVDDNNAVNEKAHGINAAVSSAK
EHWRE GRKYLNELIELNKDIADRASNINVAVDAAE
EHWRE GRKYLNELIELNKDIADRASNINVAVDAAE
PVWLA GRKQLNELIVANKQLADQATNTINVAVLA
PVWLA GRKQLNDLIVANKQLADQATNTINVAVLA
PAWKD GREHLNEVIENNRSADAATNDIVNAVNTAKGS
PIWTE GRMKLNDIITENKNVSDRATAAIDEAVLSAK
PTWYS GRMKLNDIISENKRVAQAMANDIAEAVAAAK
VTQAK YTSKVNFEFDIQDDKMSSSAQEVGESYRNA
AIQRE YRDSVKQLVNYQDDAMNTSVEAMAEVYNSTR
VVQRC YRDVAVKQLVNYQDDAMNTSVEAMAEVYNSTR
PAQN VMVLLSRLDEEVHVRVIAAASERAY AHRVIR
PLLEH RSKAIAAAVELQNCQNAAGADTLSSVALAN
PLSQK IVTAFKNIVKYQEERMELRYAELKAYGTSR
PPQRV WLABATELANFEDEMNEQAKQDAATYASVE
RLQDR YVVLVDRKMDVQAGMEHDVSEAAQGANAK
PVQAK WWALTRELKALEEKQNEEATLHAKAAYEESR
GNYSE WLKRNALIDHEEASIRVQLDNOVATASQEF
GDYSE WLRVNALIDHEEASIRSQDDVOVATASQEF
PYFTQ WLAINEFIDYQENANSGLTHQLRSDVSEFA
PAFVA WLGAINEFIDYQELNKSIGGEVRSASGFR
PAFVE WLERINEFIDYQENANQVIAERTRAVARGFAE
PAFTA WLASINAFIDLQEAKNRQAAQAVATARG
PAFTD WLASINAFIDLQEAKNREAAEAVATARG
PAFVE WLGVINEFIDHEESKNQKATPIARBTAGC
PSFAN WLKVINEFIDYQESLNQQLTPIARAEANGFQS
PAFTG WLTINEFIDYQEQYNQVLTPEARYIAGGFQS
PAFTG WLDICINQFIDYQEQYNQVLTSEARGTASGFQ
PAFTG WLDIDINQFIDYQETRNQVLTPEARNVAGGFQ
PKQVA VMARLDELQQLQENLMTASADEVSAVSTISS
ERQRS YFDVADGLTQYQAKLLAVSGKEAQTFFVSRN
DRQTA YMKGVDELQYHRAAVEKLGQGVGTAGKRAQ
KSQAD YIAAVNELIAFQASAMEKEGKNAEAMVNNAKQ
PLQNE LVEALDNMTNLQRKANEVALGK FDAYQATE
PLQCC WLALEEMAAALQERGAAMVDAADAYAN
PAMOR NODAIQCNITLQDKSAAAAADAAFRSMDSRK
PAMOC NODKTAENIALQNKIAGDASAAALQSMDSRK
PALKC NODKTAENIALQDKLAEEAAEVALESMDDSRK
PANOC NOAALDAYAARQRTLTGAACDDANTAMDHGR
PANOC NOAALDAYAARQSRGAAYDDANTAMDYGR
PATEA NOSALAEYSALQKRAKTAYEDATAMARGE
PATQA NODALLEYSALQKRAQANALATTAMARGE
PATQA NQDTLLEYSALQKRAQANALATTAMARGE
ASSDE VDTIMTDLSSLREESAASANAETGAIHSSSK
ASSDE VDNIMTDLSSLREESAASANAETGAIHSSSR
LIKDSRALEDLLKIVKRDERAKANMDESTSVN HSR
D LKDSHALEATMATMVKRRDDRARNSEARSVYTSR
QVNGA VEKONMDLVDLNGTQMEQQAQADATQVSSSR
AVLDR ADTTLTALIESERAAMTREAAQASQYAEAR
ELAQO TNAALAAHRAFNVLDQAGSNEAKDITDRA
PMFVA YDQASASVIAISLQKRAEDRQAATQSQIS
GLFTA YQCAIDALESFQVTRQKARYDAAGARFHR
ALFVA YQEAIDVLESFQVAREKARFAAGVHFHR
SLYSX YEKAMIALQQLDHBGAQRYQAAQDL

Figure B.1 continued

Xaxo_21108114_1-134	S	----	FARRD	LFAPKLVELTKFNVAHMDS	ETIAQAEATYRRSVL	----
Bfun_48787402_62-215	S	----	PLFND	MSDTNDRLSALYANAKRS	YENAEYRS	----
Bfun_48788521_31-187	S	----	GLYTA	MNASQGALENYLDQASD	ANERSAALPH	----
Rsol_17548524_27-187	S	----	YNFRT	YASQSDRLNAIQTEVSA	ALYEASQSFPG	----
Bcep_46321947_17-173	S	----	SLFTE	ASDSMDALGRICMSAR	SVTFDAAQARFH	----
Bfun_48788500_33-185	S	----	VAYND	LANADDALRKYQFTSA	KEGYDAEESSE	----
Bpse_53719442_41-196	S	----	NVFND	FSLASEALRAFOLKQAS	VNFSDAQSVYAASR	----
Bcep_46320035_33-188	S	----	VKYND	LTASEALRNQFSDAQ	RGYDHAESVYETLR	----
Bcep_46312035_33-188	S	----	ARYND	LATASEALRNQFSDAQ	GFYDHAESVYETLR	----
Rsol_17549625_31-187	S	----	ALYRP	LGDKVTALSRMQEIA	RASVYEAQREHDC	----
Reut_53762140_18-175	S	----	ELQRA	MSSAHEALQKFQITAG	QDNFDAAQARFETIR	----
Rmet_48769257_19-176	S	----	DLQRC	MSAHEKLEKQFDTG	KANFEGAQSRYETTR	----
Rsol_17428475_31-189	S	----	KSFRE	ANDASQALGKQQLTES	KANFDDSQQAYARIRN	----
Reut_53761466_20-177	S	----	KLDIT	LTAASQDLSSAQITAS	AHRVYEESSQRYN	----
Rsol_17548728_31-187	S	----	PMFVQ	LSAAVDALDRNQAEQ	AKAAYEGAVTRSON	----
Bfun_48787377_20-177	S	----	FLSLA	LTNADALDTWQKAHG	RQAFADAQRLHBR	----
Bcep_46322449_38-195	S	----	FLSVQ	LAQATDALDAYQAARG	KDVYDTAQTYNNMR	----
Bmal_53716899_85-242	S	----	FLSVQ	WTKATNALDDARAA	YGKAAVDDAEQMYGWR	----
Sone_24376324_29-192	S	----	QTEV	AGEETMNLRHENDRA	QEMVLQSENAYKTAK	----
Cvio_34104168_23-184	S	----	NEAKA	LNQALIDHIEFNKLA	DQLSKDNAAYATA	----
Rmet_48770444_25-187	S	----	ASLA	LRKTVAEHLEYNTR	KGSIJAVENNNKAQHARS	VS
Psyr_28870737_37-189	S	----	NAADC	VENTLKKLIGINDKA	ERAGNQQADDAYQOT	----
Psyr_53693339_33-189	S	----	SAADC	VETGLKKLIGINDKA	EKAGANAAYQOT	----
Pflu_48729692_32-188	S	----	PGGTV	LDKTLQMTILNQCG	ADTAAKSAAMYYQA	----
Psyr_28872674_33-190	S	----	DRASA	YQETLTTIRDENAKE	AQCSGADATSVYNH	SVN
Psyr_46188178_1-168	S	----		QALKAAAYOEKLTL	RGHNAEEAVSSGKDAT	AVYDHS
Psyr_28871675_33-190	S	----	ANANA	YQEKLTTLREQNAEE	AVSSGKDATAVYNH	SVN
Rsol_17549061_29-185	S	----	AVQST	YFDALDALVAYQTQ	LMVDTTGRALTASER	----
Reut_53760762_34-191	S	----	AAQAS	YFAPLDALMEVGKAV	SAKESAEANEAYRSK	----
Rmet_48770042_24-188	S	----	SSQAA	YFDKLDALIAQLGQ	KLAVEVEDATARYA	FTRN
Reut_46131652_18-177	S	----	AAQRE	YFEKLDEIMDGGRN	LALAAVKQADAGF	WTRN
Mmag_46203065_25-186	S	----	ANYLC	ASTAARGLIDINLA	ASRTADSEIRRAQT	DAR
RspH_8250660_26-186	S	----	ATQKK	REELAAAIVAQQL	EALDAAERDVQA	IMDEAKK
RspH_22958341_26-186	S	----	ATQKK	REELAAAIVAQQL	EALDAAESDVQA	IMDEAKK
Cvio_34103819_33-190	S	----	PANNI	LITALDEMGAFQA	CQMKDLSHEAASAE	SARN
RspH_7532754_24-187	S	----	EQWLA	METRLSALLAHHT	QQLTDAASAEACRQ	CBISRL
Pflu_48730667_20-184	S	----	KNYRV	VMDLTIINTNSND	QVSEAAKRSVLT	ESSAKTS
Psyr_46188223_34-196	S	----	SGYLA	VMDQLTTIVNSNN	RQAGEAAVSDQ	TQNSAQRN
Psyr_46187354_20-176	S	----	GFENK	ARGYLQIMIDSNK	RQIKEGAEADRLQ	STS
Psyr_28870740_34-193	S	----	DGFVK	LRGYMKTIMDSNN	RQIKEGAIAAEKLQ	SS
Psyr_53693337_34-193	S	----	SSFTX	VRSYMRVMIDSN	RQIKEGAAAAADL	KASS
Gmet_48844820_32-197	S	----	PLFKP	ANEALEKMKSEFA	AAKEDFERDDKTY	RTDR
Daro_73279524_186-351	S	----	PAYER	ASKRADDLYQLQ	ISRGKTQLEETDK	AYQBR
Bcep_46321623_26-185	S	----	QYRDE	LEGIVDTLRIEKN	RQKDDAISALNGM	LATTAT
Bfun_48782649_1-172	S	----	QYRDE	LEGIVETLVRVKN	RKDEAITTLNATL	ATT
Reut_53761418_15-174	S	----	RYRDE	MEKIVQTLRVEKN	RKDEAITALNGM	STTT
Gsul_39995862_33-195	S	----	FLYNN	PAQALAEIVETSI	KEGGEVYDADMAS	YRRS
Gmet_48846722_1-129	S	----	FLYDN	PAKIIASLV DANVK	SKAMYAEDMASYRR	ALVEMS
Ecar_50119387_32-189	S	----	KYRSQ	LMKDLAKLVEML	ELATAKNIVASAD	STYRTSQ
Rrub_48764496_1-167	S	----	FFYLA	LVKSTDRLTELSQ	IAA VAAANIDHTN	RIAR
Rrub_48764497_1-152	S	----	FFYLA	LQVSTDRLTELSQ	IAA KASANIDHTN	KVAS
Mmag_46200981_38-196	S	----	KVGRE	ANEALAQIVVAKE	KSAEQLAVASQQA	RQAVT
Cvio_34102727_35-197	S	----	PLFQA	ADDVLTLSNINI	KLADKSLDEAQ	GKYQALRT
Sone_24372094_36-192	S	----	PTFCQ	ADATLSAIVNKIV	DLGKKAYDDSD	VVVAQLGQE
Naro_48849030_29-188	S	----	DSFYA	VEDAILAAIEVNV	KAADAVSAQSEET	IYASART
Wsuc_34558217_24-182	R	----	LMSRR	LTDAIDEHITYNEN	LAEKNAILAANK	KEEA
Cace_15893832_34-189	S	----	GDSNK	TOKKLDMMKLKTN	IDEASQOKTYIK	TVTSRGE
Ecar_50120989_24-182	S	----	SEYAT	LOKIADRLIKLQ	SDVGQELYSESQ	SFFNK
Wsuc_34557239_33-189	S	----	THMDE	CEELMEQIVASNE	ENLSITDKETN	ELYEGM
Rpal_39934745_190-343	S	----	DLFRA	AKTELDKLVALQ	VCGARDEYSSG	QREY
Tden_52007871_1-136	S	----	ETYPF	VRANAELIASQL	KLAQSEFAAAQSR	FVMVRN
Sone_50261353_32-187	S	----	QVIDP	TSSTISELTALQIK	IADEKIAVDKEL	YDSS
Mmag_23013720_28-169	V	----	AKFRE	AATPLRKILDT		
Mmag_23013426_39-196	S	----	QDFRK	AATPLRKILNDY	QREANASTLEGE	ANYASDRT
Neur_30249816_32-185	S	----	PKYEA	VHGSNLNQLIGL	GESVTAQEYELQ	AMTS
Mmag_46201783_20-177	S	----	KSFAD	MTETLRALLDQVR	SVSKKEEFAAEDS	YAVSKT
Cvio_34101575_31-187	S	----	PALDT	VSQRISALVDLQ	LVAAEFSSQSQRA	AFSH
Gsul_39996476_32-189	S	----	PAIDT	VSAKFSSLVDDQ	LKIARQBYDHS	SGGLTRASRT
Pput_4235480_32-187	S	----	PLIDT	LSEGLSHLTQIQ	VEESKRAYDA	AVLVNDSRT
Pput_32469921_33-187	S	----	PLIDT	LSEGLSHLTQIQ	VEESKRTYDAA	VVLVDSRT
Naro_48850981_34-191	S	----	PRGAA	LAKALDALRDNV	KLGODASDA	AVAGATSTRS
Rmet_48770115_21-183	S	----	RAFTV	ANATLATITERN	VTGANEATQAE	AVYRHART
Xory_58582478_34-191	S	----	TLHNN	AKDSLAALIAED	NMLAQAAKTKA	EKVHATS
Xcam_21231317_34-191	S	----	TLHNN	AKDSLAALIAED	NRLAQAAKAKA	ESVHATS
Dvul_46578600_31-194	S	----	PVYNA	VATAADTVQTE	NAAAASSAMLACT	TLTVQERRTGT
Cvio_34102891_42-206	D	----	ISDC	MLETIDKHIDYNI	VIKAF	EAUVTKNA
Retl_21467290_23-187	S	----	KLYND	SGALLDQDVAVN	KRGVAAVGN	TMGAVSST
Pflu_48728799_32-190	S	----	TVVDG	BGKQLNDLADL	FARQVAAESQNS	AAHYETSRT
Psyr_28870443_33-190	S	----	QIVDG	BGKQLNDLADL	FVTRKVAEGKSA	EAAQYDKSRD
Psyr_23472455_33-190	S	----	QIVDG	BGKQLGDLAD	FYLTQVDAEGKSA	GAGQYDKSRD
Pflu_48729489_15-172	S	----	PLADE	IAVTLRELVELN	KHNANLATEAAR	LVFTNSR
Pflu_48733189_32-188	S	----	DGTDK	MGEQNLKLI	AINAADAKTAS	QAGEYYNSA
Pput_26987059_33-188	S	----	ANSEC	INKVMDTLVR	INTDOTRATNEKA	ANOYAGA
Psyr_28867696_32-188	S	----	SNSDC	MNVVLGKLVE	INTAQINQVKK	DAASREYDSA
Pflu_48732089_27-183	S	----	DNSEA	INTVLNRLMQ	INGQOISE	TNQAADQYSSA
Ctet_28211515_29-189	S	----	SAFEK	VSSSLKLLAQ	LKNRMSENDS	LEGDKRYDIAMKIN
Mmag_23011382_54-210	S	----	PRIKA	RAQALDDIVK	LNNQAADGIVA	QSEELARSAR
Mmag_23013949_31-187	S	----	QLYAC	AQKHATKAVEIN	QCGKDAASHAG	DIYDRSK
Sone_24373012_30-190	S	----	KTYNE	YSDLLQLSDINT	KGAQASNYGDEI	YDESIKMT
Paer_15596448_31-190	S	----	QRYEQ	QORTLKQILAL	DREAAASTAAE	ADVYGRAS
Paer_53727587_31-190	S	----	QRYEQ	QORTLKQILAL	DREAAASTAAE	ADVYGRAS
Paer_15596805_32-188	S	----	QYRFA	TOQQQLALID	FNWKGASQASH	TAEVYHSA
Dvul_46580384_49-207	S	----	GHEQA	AMDALYAAVE	INDKGSQAQSAE	ATNVSSAR
Psyr_28868132_32-190	S	----	LHFDZ	VTSQLQKML	ELNDAGATAAG	DKGSQDLETAR
Psyr_23472827_32-190	S	----	THFDE	VTSQLQKML	ELNEAGATIA	GNKGTSLYETS
Gsul_39995789_34-190	S	----	KLYNN	AGALIDKRLT	LNTQEAKDAS	ARGDKLYSSAR
Rrub_48764903_34-190	S	----	AAFNC	MRAALVEAID	DLNNKGA	KAAEAYQSDSDAC
Bjap_27377659_30-188	S	----	KMAQA	ADPLLQKG	IELNNGAELET	RQAADSYA
Bjap_27378042_32-190	S	----	KIGIG	SDEVLKRD	IDLNRGGDQA	ARDADSYSFA
Rpal_39937697_28-190	S	----	PLADC	MDVELQKDI	DLNNKGAD	DASSAAATYSS
Rpal_39937696_30-192	S	----	PTGSR	VDATLQKDV	DLNNKGADQ	STALAEATYSSA

Figure B.1 continued


```

Rpal_39934710_32-186      PISAC  TDEVLRKKDIDLNRGAAGATASAEAT
Psysr_28871756_32-186     PLTNC  METQVEELTQFNNKGAALAGVQATQVYDSG
Psysr_46189040_15-171     PVTGE  LQTAIDALVQMNIDQADQLSVRAEAAYES
Psysr_56459719_31-188     PLTDK  ITQQLDRMYGAERSLINQAADNAHNVAENS
Bbac_42522500_32-194      KAAKV  YDDAMTKLVAFQRASGVQVYVKEARSTTRTT GS
Bbac_42524709_29-190      KAAKV  TYAPLINETTYQVQYADRSHEATAAEKQAR
Rmet_48772490_1-153       ALNQC  LDLTIAS VDLVAMAAALRQES TMLS
Rsol_17549248_31-188      ATGTS  TSEWLAQDIEINLEQAKDMDTDSAAITYGR
Reut_53761145_32-187      AAGDV  VGEILTRGVEVNLQAGDYHRRSLTEARS
Linn_16799806_33-190      TKFDD  ATKQLNVLNEMNYNEVEKSSQAISDFGVK
Lmon_16802765_33-190      TKFDD  ATKQLNVLNEMNYKEVENSQAISDFGVK
Lmon_46906974_33-190      TKFDD  ATKQLNVLNEMNYKEVENSQAISDFGVK
Dhaf_53686096_14-173      LQAAK  VEGILQDNIQESLAYNEELQRAESQDFAKA
Bsub_16080422_25-184      ASFED  MKKTIITQLVDLNLQEGSNTAVKETKAVYHKG
Ctet_28210899_35-191      EVRDI  MFSQLNELIDLNIKMSREAKAVSTSTYDKSEK
Cace_15896748_32-189      GYRAR  INDLLDEEITYNQKVAKSKYEDSKQYKKA
Cace_15896020_28-189      HYSDI  MNKLNLYIDYKSDITLYNYGDIKGOYKSA
Cace_15896019_33-189      PYRNA  ALQKLDADIKYNKAVKANDYSSSKVOYKSA
Cace_15894666_33-188      SYRNT  MFGYLDKEIALNTKLAKSDYDNSKVYQNA
Cace_15896638_34-188      AGDI  LDKSLTALRDYETKKAQVVMGSSDSIYSSNK
Ctet_28211181_69-221      KFGNE  AETALDKLQYDIKLABEDKIESDIYIINRN
Ctet_28211184_33-183      TGDK  TQASLDNLEKYDVKLSEENKIESDKIYITNRN
Dgig_4235392_33-193      PLGVV  PDQLLSDMAEENRNEAEKAAQAAGVYLEARE
Esp_46114138_32-187      KTRDE  VANTAKSLMMSMTKQSKDINALNQEQKTA
Bcer_52144881_34-189      PNLSA  TVGALKNLSDYNEKLAEGLYNDSSEHSYKQ
Bthu_49476793_35-189      PNLSA  TVGALKNLSDYNEKLAEGLYNDSSEHSYKQ
Bant_47525640_35-189      HNLSA  TVGALKNLSDYNEKLAEGLYNDSSEHSYKQ
Dhaf_23121705_34-189      HLSDC  LVEDMRALGEYYASLSLQASLDYEA5FKH
Bagr_29170613_36-187      KTL5W  INQEQLEWANYKSSQADRMQOQNWDDSKK
Dhaf_53685321_20-174      SOVDA  VNDLLKELADYNAQLADEEEIKSQAAMTSK
Dhaf_53685000_1-147       DDL5W  VNALLLEELADYNAQIADDEGIKSKSLASMVSK
Blic_52002118_32-188      PORET  VNQLIEDIQTINADNAKTIYQDSKEAGS
Bcer_52142147_38-193      PNMOE  SIQSIRELILYNSNDAELLQKENNNGAONT
Bcer_30021489_34-188      PNMOE  SIKSIRELILYNSNDAELLQKENNNGAONT
Cace_15893416_33-189      VAHDK  LATSIGNVTDMKKELAKESTONSATAKTS
Mlot_13488383_33-191      PIFVK  LQSAIQAMVADHRGANSATORISSIVRTAE
Cglu_41326927_358-461    TSFDE  LDTALAEIADSRSSMSRYIQSLQATE
Xcam_21231331_61-200      VAGDT  HLLNTELSLQSSVASV ST
Xaxo_21108113_80-225      SGGDT  HVLNTELSLQSSVASV VAG
Rleg_15072891_30-186      PQAEI  VNKAVADLVAFILSQAEGVFAASGASQSA
Rleg_2665910_26-185      PQAEI  VNKAVADLVTFILSQAECFVAASGASQSA
Vfis_59712649_32-186      GYYSR  ALTSLLDDTLALNDVSVNQISNDVQQAES
Wsuc_34557328_34-190      PQAGK  TIKGVQDLFKDKLEDAKLSDSNDELAASQC
Wsuc_34558241_35-189      VQGRK  CGDLLNKFMEFNVKLAETISAENDQVALETS
Cthe_48860112_13-169      DIGDT  LRDYFEAFVEYNTAAKEKVDENKQVASTAST
Cthe_477735_1-125         DIGDT  LRDYFEAFVEYNTAAKEKVDENKQVASTAST
Vpar_28900346_32-187      FETG  LESAIDRLVELNQTIVQEDIASAHEAVRSA
Vfis_59713352_32-189      DFNG  VGDAMDGLVDINLGFVSNRNTSISMSSVST
Vfis_59713351_32-191      LFEN  VGAALDGLVQINIGFVANNRTAMHSVDNVTATK
Vfis_59713353_32-185      EF5FK  IETEVSALTBIKQAMNSNRVQLDSISRLSQMS
Ppro_54303603_31-193      DFEA  VDSDLNELIRLLKVAMDSNKNHILSSVNLSSSS
Vvul_27358478_31-191      STFES  IETEVNKLWMLKGAMDSNKNHILSSVNLNTTA
Vcho_9655802_32-191      FFEA  LESDFTLIGILHQAMDSNKNVQLSSVKTLNST
Dhaf_23113288_34-189      KLSAN  FSAAIDNLSLIKEQTGHEVAAANKARAESA
Gsul_39997674_35-189      KAALR  VQELLSKLVDAKQAQAKLTAERNEHVASTS
Dvul_46580722_34-190      GIARS  VDDSIQKLAOLKIDLAKKSDANTAAAK
Gsul_39996400_29-188      AYERK  IDDAIKKIFDMKIALAERNGLNSATAHSAT
Gmet_48847252_32-190      FEKK  IDESIKKLFDMKIAGAKKKKMNAAAQSA
Gsul_39998033_33-189

```

Bagr, *Brevibacillus agri*; Bant, *Bacillus anthracis*; Bbac, *Bdellovibrio bacteriovorus*; Bcep, *Burkholderia cepacia*; Bcer, *Bacillus cereus*; Bfun, *Burkholderia fungorum*; Bjap, *Bradyrhizobium japonicum*; Blic, *Bacillus licheniformis*; Bmal, *Burkholderia mallei*; Bpse, *Burkholderia pseudomallei*; Bsub, *Bacillus subtilis*; Bthu, *Bacillus thuringiensis*; Cace, *Clostridium acetobutylicum*; Caur, *Chloroflexus aurantiacus*; Cglu, *Corynebacterium glutamicum*; Ctet, *Clostridium tetani*; Cthe, *Clostridium thermocellum*; Cvio, *Chromobacterium violaceum*; Daro, *Dechloromonas aromatica*; Dgig, *Desulfovibrio gigas*; Dhaf, *Desulfitobacterium hafniense*; Dvul, *Desulfovibrio vulgaris*; Ecar, *Erwinia carotovora*; Ecol, *Escherichia coli*; Esp, *Exiguobacterium* sp; Gmet, *Geobacter metallireducens*; Gsul, *Geobacter sulfurreducens*; Hhep, *Helicobacter hepaticus*; Ilol, *Idiomarina loihiensis*; Linn, *Listeria innocua*; Lint, *Leptospira interrogans*; Lmon, *Listeria monocytogenes*; Mdeg, *Microbulbifer degradans*; Mlot, *Mesorhizobium loti*; Mmag, *Magnetospirillum magnetotacticum*; Msp, *Magnetococcus* sp; Naro, *Novosphingobium aromaticivorans*; Neur, *Nitrosomonas europaea*; Npun, *Nostoc punctiforme*; Paer, *Pseudomonas aeruginosa*; Pflu, *Pseudomonas fluorescens*; Ppro, *Photobacterium profundum*; Pput, *Pseudomonas putida*; Pres, *Pseudomonas resinovorans*; Psyr, *Pseudomonas syringae*; Raqu, *Rahnella aquatilis*; Retl, *Rhizobium etli*; Reut, *Ralstonia eutropha*; Rgel, *Rubrivivax gelatinosus*; Rleg, *Rhizobium leguminosarum*; Rmet, *Ralstonia metallidurans*; Rpal, *Rhodopseudomonas palustris*; Rrub, *Rhodospirillum rubrum*; Rsol, *Ralstonia solanacearum*; Rsph, *Rhodobacter sphaeroides*; Sent, *Salmonella enterica*; Sfle, *Shigella flexneri*; Sone, *Shewanella oneidensis*; Styp, *Salmonella typhimurium*; Tden, *Thiobacillus denitrificans*; Vcho, *Vibrio cholerae*; Vfis, *Vibrio fischeri*; Vpar, *Vibrio parahaemolyticus*; Vvul, *Vibrio vulnificus*; Wsuc, *Wolinella succinogenes*; Xaxo, *Xanthomonas axonopodis*; Xcam, *Xanthomonas campestris*; Xory, *Xanthomonas oryzae*; Zmob, *Zymomonas mobilis*

Figure B.1 continued

[illegible]

Figure B.2 Edited seed 4HB MCP alignment with VISSA visualization.

Figure B.2 continued

Ecol_2506837_36-190	PLFEMVATSRINDENKYNATTELITLIDLV	-G	IGAYFAQPTQCMNAMSEAFAYVALSSBKRYLDVTDNADDTRE
Ecar_50120625_36-192	ARQDDISRGVKNFVVALNAELITLQIN	-G	SKFPIEQPTQDNFBEKAYVYKAENDKIQAGIKANDDAYDS
Styp_16423100_36-192	PQSEAAFLIKRTYDITGHGALAEILQLG	-G	INEFFDPTQSYQDAQFEQKYVAYMGNDRLDIADVNSSNYSQ
Sent_56416317_36-192	PQSEAAFLIKRTYDITGHGALAEILQLG	-G	INEFFDPTQSYQDAQFEQKYVAYMGNDRLDIADVNSSNYSQ
Ecol_43218_36-192	PTSTAAAEIKRNYDINHNAELIQLG	-G	STSSLISRPDRINGFEQKYVAYMGNDRLDIADVNSSNYSQ
Ecol_16132176_36-192	PTSTAAAEIKRNYDINHNAELIQLG	-G	INEFFDPTQSYQDGFEQKYVAYMGNDRLDIADVNSSNYSQ
Ecol_26251236_36-192	PTSTAAAEIKRNYDINHNAELIQLG	-G	INEFFDPTQSYQDGFEQKYVAYMGNDRLDIADVNSSNYSQ
Sfle_24115585_36-192	PTSTAAAEIKRNYDINHNAELIQLG	-G	INEFFDPTQSYQDGFEQKYVAYMGNDRLDIADVNSSNYSQ
Ecol_13364793_36-192	PTSTAAAEIKRNYDINHNAELIQLG	-G	INEFFDPTQSYQDGFEQKYVAYMGNDRLDIADVNSSNYSQ
Bjap_27376721_106-258	PLFDDLEQFEKQKRVQVQIDFLPLVLE	-G	ISPAAREWDNQINRIVRSKLADLEALQSYSDKRAEADQAD
Cace_15896003_35-189	KANESQQTIKETSNEGLTVQVLEAAQ	-K	CPAANLNKSNSQNRQIKFEETKIKVDNNTOSRSENANATANV
Msp_48832885_28-179	CGQETAERLSRADFATETITVDFAR	-E	SRR DYFA R LPRFQETKSLANALLEMSQNSMSEANDRARAKA LT
Gmet_48846841_26-183	DREKQGTAAIQENQYQDGFQGLMQV	-S	ITSTQANRA MKPVAE KATKLEITEGNRAVEMSKDTKEVDAQ
Bbac_42522983_19-180	TEPMERFAEQKQVABRYEKLLIMVIBISTG	-G	FALASQLQGLTWIGTITGHHSSSEMRYVQAADAKRAEDATRIN
Bbac_18073058_46-207	TEPDRAVYQAKPLFPFYVYASMEKVALIRSD	-AK	VKEABALLDGRNIEIGKAVRWNGAVITLAKYAKMDGVTEANSTAYK
Bbac_42522982_46-207	TEPDRAVYQAKPLFPFYVYASMEKVALIRSD	-AK	VKEABVLDGRNIEIGKAVRWNGAVITLAKYAKMDGVTEANSTAYK
Ilol_56461443_33-191	SATERQWYDQVKQNETWTATAKQVITLHN	-Q	TEAARLSQSGQGFQPEAMRNVLDEQRKEKALQSQDIDIVESQSN
Mdeg_48864484_35-190	KTEKQVGVDFNDRPKRRATTERLSSLD	-S	RASAVQISMGAAQTQDEMDRIDTTLTHRKSVLSDLEATAASAAN

Figure B.2 continued

Figure B.2 continued

Rpal_39937696_32-192	TPEERTLYEEFSKELNVYLDGAKKAVEMSKSV	-GQ----	STAELEGYLTKTLGPIGSRVDAILQKQDVLNKKGDAQSTALAEATYSSA
Rpal_39934710_32-186	TPEERKLYDSLVSASWTNNTYSLAPTGLEMSRKAA	-GE-----	LPAANDYMAKTMGPISAQIDEVLKRDIDLNDKGAAGATASAEAA
Psyr_28871756_32-186	SDQERQLFLAVKKSYDDYASQLDLLLEPLLK-A	-GD-----	TASIVKLVATGIRPLTNQMETQVEELQFNNKNGAAGLVQATQVVDG
Psyr_46189040_15-171	AADEASLYQTVTASAQKFASLLDTVLELIQ-K	-GA-----	NADAVTFTDNIIVPTVGELQTAIDALVQMNIDQADOLSVRAEAAYES
Ilol_56459719_31-188	DFOACKLISEYEELDKKEYWKLEQNFKNLIV-A	-GL-----	VDAIASIRENSILPLTDKTIQOOLDKMYGAERSLNQADNAHNHVAENS
Bbac_42522500_32-189	KKERELYEKVDAAMLAFKDLGGKVVQVYSR-T	-GK-----	FEDRAKMLEIFLYDOPKAAKVDDAMTKVAFQRAQSGVQVVKEARSTTB
Bbac_42524709_32-189	SAKDRASYKEFTIAGWHEFKSFSGGEILSMSANY	-GQ-----	NESKVVSLIRCPAKAQKIYAPLNNETTYQVQVADRSAREATAAEAKQA
Rmet_48772490_1-153	NOHECVLATHGRLTLFSGTVRSSLERIS-	-GD-----	HDDTRDWLEQNVNALNOLDTLIASVDLVANAAARLRSES-TLMS-
Rsol_17549248_31-188	SKKEREVATKINDTLPKFNDAANKAVDALR-A	-GN-----	LDAAGQIINGNSATGTSISEMLAQDIEINLEQAKDMDTDSAAVYQS
Reut_53761145_32-187	SPREDMADQDLTALARFADLVAQELHLLE-	-GD-----	RAAAARLQESDLGAAGDKVGELITRGVEYNLQAGDYHRRSLTEARS
Linn_16799806_33-190	DKKQLAYFEELADMKSAASSVISDTSSALDD	-AE-----	LLGAQNRYYQNVKTKFDDATQQLNVLMNMYNEVEKSSQAISDFGVH
Lmon_16802765_33-190	DKKQLAYFEELADMKSAASSVISDTSSALDD	-AE-----	LLGAQNRYYQNVKTKFDDATQQLNVLMNMYNEVEKSSQAISDFGVH
Lmon_46906974_33-190	DKKQLAYFEELADMKSAASSVISDTSSALDD	-AE-----	LLGAQNRYYQNVKTKFDDATQQLNVLMNMYNEVEKSSQAISDFGVH
Dhaf_53686096_14-173	TDEEKELLGEFKREAAVYRDLAKKALTAVA	-I-K-----	VALAGDFNQAGLQAKVEGIDQNIIECSLAYNEELQASEQOFAKA
Bsub_16080422_28-184	TDKEQKLFEOQLTKVNTYMDIHAQIIESGR-T	-ND-----	MDKARGLLVQTEASFEDMKKTIITQLVDLQBGSSNTAVKETKAVYHKG
Ctet_28210899_35-191	TRENGIYNLRKSSSDYRSIHNEIMNAVH-K	-ND-----	YKRAKPHVNOISEVRDIMFSQLNELIDLNIKMSREAKAVSTSTYDRSEK
Cace_15896748_33-189	DEKDKKIPTQYLENLNGYNNAKKFFETAR-	-G-----	FNAIKNQPSIFDGYRAKINDLDEBITYNOKVAKSKYEDSOKYKKA
Cace_15896020_32-189	MEKDKELFSEFVNLLKKWASRNKNVISYVV-	-GE-----	YAAKTEFSDTTSHYSDIMLNKLNLTYDSKDTITYNYGDIKGGYKSA
Cace_15896019_33-189	TDEERTLPYQFLDNLAQRQAKREKVIAYAR-A	-GD-----	YMSASVEFKTHPTPYRNAQLKLDATRYNAPKAVVENSASSKQVYKSA
Cace_15894666_33-188	IADRKFLFNQVNIELKKWGIARSEFVDAVN-	-GD-----	YDKANEYTKASSYRNMTFGYLDKALINALNKLAKSDYDNKSVVYQNA
Cace_15896638_34-188	DSNENASISFSNSNYEYVNLWTELSALKTG-	-GK-----	LAETKINRFARLGDIDKSLTALRYDEYTRKAQVVMGSSDSIYSSNK
Ctet_28211181_69-220	DDEIKYLEQFLQCYKNYRDLSEKELFKHLBR	-GE-----	KLTREQNDKRFKFGNEAETALDKLQEVYDIKLAEDDKIESDBIYIINR
Ctet_28211184_33-183	DKTQAKYIKAFTESYNDIYMTKEYLERIKK-	-GE-----	KITQKDNNEILKIGDKIQASINDLSEKDYVKLSEENKIESDKIYTTNRN
Dgig_4235392_33-191	TGKERELYDTFKPAYDAYLAENAKLRQMMNK-	-ED-----	KDVTDFVLNRVAPLGVLPDQLLSDMAEENRNEAEAAQAQGVLYLAR
Esp_46114138_32-187	TQSQEKQVENLKDVSLAYDNGSLSTIQDLIVT	-GD-----	KEQGYRFYSESLEKTRDEVANTAKSLMMSMTKQSKDINALNQOEKQTA
Bcer_52144881_34-189	TGEEKKLLSQYKEKLENYKKHRAQMLDFIH-E	-NK-----	LNQAYSFYLTTVFENLSATVGALKNLSDYNEKLAEGLLNDSHSHYKQ
Bthu_49476793_35-189	TGEEKKLLSQYKEKLENYKKHRAQMLDFIH-E	-NK-----	LNQAYSFYLTTVFENLSATVGALKNLSDYNEKLAEGLLNDSHSHYKQ
Bant_47525640_35-189	TGEEKKLLSQYKEKLENYKKHRAQMLDFIH-E	-NK-----	LNQAYSFYLTTVFENLSATVGALKNLSDYNEKLAEGLLNDSHSHYKQ
Dhaf_23121705_34-189	DPQSQELFLKIKASYQKYDQVKNQIVAFAR-	-NL-----	NDKAYNLBYTSGGHLSDQLEDMMRLALSDYNEKLAEGLLNDSHSHYKQ
Bagr_29170613_36-187	DEKEQERVSKLKELEQYSKDQNEVVOQLIH-E	-NK-----	TDQAYAYRRAT-KTLSWINGQELQWANYSSQADRMQOQNWDDSKH
Dhaf_53685321_20-173	DAHEQEQAFLMDTLQLYRTERDKAIDMAT-A	-GN-----	ODEAFRYFSAHAASQVDVANDLLKELADNQADEEBIKSQAKAAITS
Dhaf_53685000_1-146	DTYEQERLPHLMDELQYRTERSKAVIDLAI-A	-GK-----	ODEAYTYFAANAADDLVVNALLLEELADYNQATADDEGKSKSLAMVS
Blic_52002118_32-188	DNVSEKYSEFKSEYKKLODISSOMLSIAV-K	-NE-----	NKADYDYLKEMEPORETVNOIIDEQTLNADNAKTIYORDSKEAGS-
Bcer_52142147_38-193	STKEKELVNTFHETFNELATQMRKAQELGH-G	-N-----	NEEAYAYYLKEIEPNMKQSIGSIRELLLYNSNDAELLQKNNNGAONT
Bcer_30021489_34-188	STKEKELVNTFYETFNKLATQMKKAQELGH-G	-N-----	NEEAYAYYLKEIEPNMKQSIGSIRELLLYNSNDAELLQKNNNGAONT
Cace_15893416_33-189	TADGKTLISIKIDAKKEYVQYQAQAMELSH-	-NK-----	NVEAMAIVRNOLTVAHDKLATSINKVNDMMKELAKESDQNSATKTS
Mlot_13488383_33-191	SIDKONYAVVKELANQFLAETAVLDTSSAALS	-GE-----	FNESFAIKNHYEPIFYKLQSLAQAMVADRGEANSATQRISSTIVRTAB
Cglu_41236927_358-461	TPEHQQILTAIRNALAANTADHDEETVLIA-	-G-----	YNGAVNAVNLKDEECQTSFDELDTALAEIADRSRSMRSYIOSGLOATE
Xcam_21231331_61-200	QEQAQARAVOQQLGOYLQAHROANRALH-D	-GB-----	LHAAQALS-GH-----YAGDT-HLMTLEQSLQOSVASV-ST
Xaxo_21108113_81-225	QEQARKEALVQQLKRYLAHQROANRALH-E	-GB-----	LQYAQALS-GH-----SGDTHLMTLEQSLQOSVASV-AQ
Rleg_15072891_30-186	TERGRELINLIKPELAKYRALSSOMTAFEN-	-GK-----	TPEATRLFKENMEPOAELVNKAVADLVAFILISQAEQFVASGDSAQSA
Rleg_2665910_29-185	TERGRELINQMKPELAKYRALAQMIALEN-	-GK-----	TPEATRLFKENMEPOAELVNKAVADLVAFILISQAEQFVASGDSAQSA
Vfis_59712649_32-186	SSEERASFQFKETWNOVYKSTHYSNNLLS-Q	-GN-----	STQANEVVLSSFGTYSKALTSDDTLALNDVSNVQISNDVQOAGSS
Wsuc_545307328_34-190	DSKDKIEIDELIANFKEYKKITILEVLKLLR-G	-GR-----	TTEADKMTSTVMVPOAGKTIKGVQDLFKDLSDSNDBLAAQSG
Wsuc_34558241_35-189	TPEDKALFEBIVREFTYKSLSLKRFELLE-A	-GK-----	TTEAKTFASQVAVQGRKCKDLNRMFENFVNLBATISSAENDQVALETR
Cthe_48860112_13-168	NETDRIQLQELKALWEKYKSLVDKEVELVR-S	-GK-----	TTEARQLLLSDIDDIGTLRDYFEAFVEYNTTAAKEKVDENKQVASTAS
Cthe_477735_1-124	NETDRIQLQELKALWEKYKSLVDKEVELVR-S	-GK-----	TTEARQLLLSDIDDIGTLRDYFEAFVEYNTTAAKEKVDENKQVASTAS
Vpar_28900346_32-187	DQRDRDAFNKVSASWVKYSAPNNEYAKLLI-N	-N-----	TDEANKTLLNGFSFTTQLSDAIRDLVELQNTYQVEDIASHAHQVASTAS
Vfis_59713352_32-189	GAEERRVFDVAVLSWKAYSNLFDRDFQLIH-D	-RE-----	LDKAHAILVNSLDDFNGVGDAMDGLVDINLGFVSNNRSTSIMSVDVSTV
Vfis_59713351_32-188	TEHERRIFDRVKNWSKTYTTLQDKDFNSLVS-V	-QE-----	TKQAQKELVNSFILFENVGALLDGLVDTNIGFVANNRNTAIMHSDVNTVT
Vfis_59713353_32-185	SADERRVFERVANSWQKYLTLQKDFNGLIVA-N	-KK-----	AKLAQBELVDSFSQFEVGTAMDGLQLANLGFVSNNRKSGIMSQVNN
Ppro_54303603_31-190	PGEESYKRLMSGWQRYINNIDNFNSAMI-L	-NN-----	KRAARSILLTDSYPSFKIETEVESALTDILKQAMNSNRVQILDSISRLS
Vpar_28900855_31-189	PGEEOQTFRQLMRQKQYLVMTDQYNESMI-A	-GN-----	KTEALAVLSNSLNDFEAVDSDLNELIRLLKVAMDSNKNHLLSSVNGLSS
Vvul_27358478_31-189	PGEEOQTYNRLMSLWSGYLSTMDKFNDAI-A	-GI-----	KDAAYPILTNSLSTFESIEETEYNKLVMLKQAMDSNKNQILSSVINGLNT
Vcho_9655802_32-189	PGEEOQTFKRLMGNWNAYTAVTDQFNQTLI-T	-QG-----	ADDAYPILANSLSFTEALESDFTLILGILHQAMDSNKNQILSSVKTLSN
Dhaf_23113288_34-189	TEEETLQFTYLDNVLENFDYHLDQVMELCH-E	-GN-----	KTFAYTVLAQDGPKLSANFSAADINLSLIKEQTGHEVAANKARAEASA
Gsul_39997674_35-189	TDEGRTLFNEYKEARKVYGGYIDNIMQLQNS-A	-KK-----	VTEAKALLHGDAKKAALHYQELLSLKLVDAKQAQAKLTAEERNEHVASTS
Dvul_46580722_34-190	SESERRDFDTFEKTRQTYQALMDKVFLLO-S	-GQ-----	HDEAVSIVNGEGROLAHRYQELLKLVQAKVSAARQSTDNSOLDADKAT
Gsul_39996400_32-188	SEETKKAYADIKAEAKAFAPHLDKIVALAN-D	-GK-----	NDAQVAYMRSDSVAGIARSVDSDIOKLADLKITDLAKKSDSANTAAAK
Gmet_48847252_32-190	TDEGKKQFATRLTLIKEGPGVRDEITAAAL-A	-GD-----	RETALDVMRSEGLAYERKIDDAIKKIFDKMIALAEERRNGLNSATASHAT
Gsul_39998033_33-189	SKDIRQEPDALRNTIAEYAPVREEIVTATI-D	-GD-----	RETALALMRSQGLAFKKIDESIKKLFDMKIAGAKKRKDMNAAAQSA-

Bagr, *Brevibacillus agri*; Bant, *Bacillus anthracis*; Bbac, *Bdellovibrio bacteriovorus*; Bcep, *Burkholderia cepacia*; Bcer, *Bacillus cereus*; Bfun, *Burkholderia fungorum*; Bjap, *Bradyrhizobium japonicum*; Blic, *Bacillus licheniformis*; Bmal, *Burkholderia mallei*; Bpse, *Burkholderia pseudomallei*; Bsub, *Bacillus subtilis*; Bthu, *Bacillus thuringiensis*; Cace, *Clostridium acetobutylicum*; Caur, *Chloroflexus aurantiacus*; Cglu, *Corynebacterium glutamicum*; Ctet, *Clostridium tetani*; Cthe, *Clostridium thermocellum*; Cvio, *Chromobacterium violaceum*; Daro, *Dechloromonas aromatica*; Dgig, *Desulfovibrio gigas*; Dhaf, *Desulfotobacterium hafnienae*; Dvul, *Desulfovibrio vulgaris*; Ecar, *Erwinia carotovora*; Ecol, *Escherichia coli*; Esp, *Exiguobacterium* sp; Gmet, *Geobacter metallireducens*; Gsul, *Geobacter sulfurreducens*; Hhpe, *Helicobacter hepaticus*; Ilol, *Idiomarina loihiensis*; Linn, *Listeria innocua*; Lint, *Leptospira interrogans*; Lmon, *Listeria monocytogenes*; Mdeg, *Microbulbifer degradans*; Mlot, *Mesorhizobium loti*; Mmag, *Magnetospirillum magnetotacticum*; Msp, *Magnetococcus* sp; Naro, *Novosphingobium aromaticivorans*; Neur, *Nitrosomonas europaea*; Npun, *Nostoc punctiforme*; Paer, *Pseudomonas aeruginosa*; Pflu, *Pseudomonas fluorescens*; Ppro, *Photobacterium profundum*; Pput, *Pseudomonas putida*; Pres, *Pseudomonas resinovorans*; Psyr, *Pseudomonas syringae*; Raqu, *Rahnella aquatilis*; Retl, *Rhizobium etli*; Reut, *Ralstonia eutropha*; Rgel, *Rubrivivax gelatinosus*; Rleg, *Rhizobium leguminosarum*; Rmet, *Ralstonia metallidurans*; Rpal, *Rhodopseudomonas palustris*; Rrub, *Rhodospirillum rubrum*; Rsol, *Ralstonia solanacearum*; Rspsh, *Rhodobacter sphaeroides*; Sent, *Salmonella enterica*; Sfle, *Shigella flexneri*; Sone, *Shewanella oneidensis*; Styp, *Salmonella typhimurium*; Tden, *Thiobacillus denitrificans*; Vcho, *Vibrio cholerae*; Vfis, *Vibrio fischeri*; Vpar, *Vibrio parahaemolyticus*; Vvul, *Vibrio vulnificus*; Wsuc, *Wolinella succinogenes*; Xaxo, *Xanthomonas axonopodis*; Xcam, *Xanthomonas campestris*; Xory, *Xanthomonas oryzae*; Zmob, *Zymomonas mobilis*

Figure B.2 continued

Figure B.3 Complete 4HB MCP alignment with VISSA visualization.

Figure B.3 continued

Figure B.3 continued

Figure B.3 continued

Rsol_17427298_25-180
Reut_53761798_27-182
Rmet_48770512_28-183
Rgel_47574299_34-184
Psp_54032248_52-203
Psp_54032135_25-175
Bbac_42524788_28-184
RspH_44522083_29-182
Naro_48850714_37-186
Drad_15807871_51-199
Rpal_39935500_31-186
Bjap_27379400_31-186
Bjap_27379387_1-124
Pflu_48731493_44-184
Psysr_46188881_30-168
Psysr_28869899_43-182
Pput_24985438_43-179
Bfun_48785827_44-182
Mfla_45522083_45-187
Ppro_54308182_32-133
Ppro_54302047_35-187
Xory_58582467_55-214
Xaxo_21108104_4-133
Psysr_23468728_25-194
Bcep_46321500_33-168
Rrub_46766644_32-188
Mmag_46204782_17-166
Dpsv_50875181_214-367
Dpsv_54308182_175-296
Lint_24214048_27-178
Lint_45658225_27-178
Bjap_27380803_69-216
Dhaf_23116533_73-221
Xaxo_21106794_50-198
Psysr_58583559_50-198
Xcam_21232865_34-184
Psysr_23468998_40-186
Psysr_28869932_13-157
Xcam_21107031_39-186
Xcam_21230299_39-186
Mfla_46121047_48-205
Vvul_27339643_30-187
Vvul_37679272_30-187
Nfar_54022117_37-190
Scoe_6562854_30-186
Mavi_41409488_37-192
Tfue_48835308_48-200
RxyL_45548044_36-189
Gvio_37520585_36-203
Ppro_54302045_49-202
Ccre_16127375_29-179
Sthe_51892221_47-198
Sthe_51892220_31-185
Sthe_51892086_31-179
Bjap_27377458_27-183
Bjap_27377456_27-182
Rpal_39933216_26-159
Rrub_48764325_9-127
Rsp_36985891_29-192
Bjap_27377617_93-243
Bjap_27377618_16-163
Bhal_10173491_26-176
Bhal_27634002_13-132
Sthe_51891861_42-195
Vpar_28900417_195-324
Ppro_54303525_181-310
Msp_48832744_193-326
Bjap_27375494_35-174
Mmag_23014424_13-172
Mmag_46201214_26-180
Mmag_23014424_180-329
Bjap_27375494_184-330
Daro_41723787_35-182
Tden_52007874_16-169
Wsuc_34557595_31-176
Mmag_23014940_31-186
Mmag_23016728_31-186
Asp_56475846_33-186
Gmet_48846957_29-184
Mmag_23014940_31-187
Cthe_48860417_33-191
Wsuc_34557782_13-175
Bcer_30019163_32-181
Bcer_47564964_33-181
Bthu_49480178_32-181
Bant_47562675_33-181
Bcer_52144326_33-181
Bcer_42780166_32-181
Psysr_38257056_31-186
Mfla_53759318_40-184
Cvio_34102321_36-193
Daro_41725375_11-174
Ppro_28900417_42-178
Ppro_54303525_28-166
Bhal_10176488_27-178
Msp_48832744_44-170
Msp_48833776_48-185
Ppro_54309570_33-174
Vfis_59711740_36-186
Vcho_9656431_37-179

SETGNIRLSQGYRQVLSQQAETAINEELAEVLNABAGQGRFL--TGK-----DEYLDPNY-KAIPHIHALMAEIRDHYAND--PEAL
SETGNVRLRESYNEVRSQRVQTQALALNGELVNABAGQGRFL--TGK-----DSYLDPNY-QALPRINELMAHRAHYAD--PEAL
SETGNQRLREGYIEAIRSQQLQNDLGLDIAELVNABAGQGRFL--TGK-----DSYLDPNY-KALPRINELMSRIHQHYAND--PEAL
RSALAIQETNAADVARIHVGRVLRVMYVDAETQGRGYLL--TGK-----EAYLTPFY-NAVTOIKELLSPLSRVYATH--PAOL
-----SKSITHAMDDLSEAHVTRTHLQRLRLDVLDAETQGRGYLL--TAD-----FAYLEPFY-PAIADLSQOTQDQVLNLEANK--PEOS
-----ESTAAANIEEAQQTREAINQLQRLMDAETQGRGYLL--TGD-----FRYLEPNY-PAIDISONLNDVLRLQFEPY--RNEI
KFLALKEYRETIALYEKTEALISMIQNVKSTLIDLBTQGRGYLL--T-N-----EQELKPYL-DANDAIBKEFELEKSAB--ATLKA
FVLSIRADRTFQAIVEERATRHRHAADLLSALQDIBIGRGYLL--LD-----PTLEPFY-AALARAPPELTALESNE--TSAGI
-----TRIQREQAARTSGILASLDEITRATMNGHT-QRGGYLL--TSD-----TRVLPFY-EGQERYSIETIRLEKMG--GELFEQ
-----NLSAESARFAQTSQGRLLTLRLLLDISNMNGHGRGYV--TGO-----PSLEPFY-EGQONFAEHLRRYEPITV--TDQR
SVYLWQKAREDNAWVAHTLEVQNTQIALAQIQRRAESAERGYYL--T-Q-----DDLIAPFY-PAASDVVPRLKKRLDVA-N--PAOV
SVYLVNKAREDSKSVVHTIEVENQINTLLLEVRAESSARGFL--TQG-----PDEQSDHE-KAAVAITPALCKLTBQIG-N--PAOR
-----MRRAESARGVTELYSPAPTEFQVHAQIAPALADKRG-----V-DNPDOVTLLEG--TEF
VLCQWEHTDRVINNAEAVKLVLDLS-GMRGYLL--SGD-----EHLDPFY-KAPRIAVANTLEILT-DN--PAOL
AICQWEHTDRVINNTRSMKLSIDMETGMRGYLL--TGD-----QHFLDPFY-KAPPLLLAELOGLKLVLDN--PTOL
NAICQWEHTDRVINNTRSMKLSIDMET-GRGYLL--TGD-----QHFLDPFY-KAPPLVLAELNGLOLVLDN--PTOL
SAMQWVHTDRVIGNANETKSIDMETGMRGYLL--TGD-----ERFLDPFY-KAPRILGSLKSLRSMVDN--POOV
-----TMNWAHESERVIGQANEMRLAVDRSMGMRGYLL--TGD-----ESFLTPFY-SGGPRFKTOIALQGLVLDN--PPOV
-----SMNWEHESERVIGKAEVSKLAVDRSMGMRGYLL--TGD-----NLFLOPFY-DSKPTQIATISGLALLVKXS--PAOT
NTATMSTAKVHEHTYVIGVNSNGLVNAVDDE-GLRGFY--GGO-----DDYLEPFY-SGKEKFOEHLKAKHLLSDN--PAOO
STERAKTSNEWVEHTYVVDIKIDYTLVNVME-GY-GFM--TNN-----SEPLEPFY-EGQEAIVKQSLTLITITADN--KTOT
GVRGQSELSDAVAINDHTYQVIATGRAMITATVNVETARGFL--SGO-----SAHLTPFY-KGKAQFEEFTRAKTIT-DN--PVQO
-----ATVNIETARGFML--SGO-----SHLTPFY-NGKVQFQEFNTAKTITADN--PVOC
LFLFIAVLVASAYRGFEVSEATNSNVHTYVLSAEQALQALINLET-GRG--FV-TASK-DAFLEPLVAGEKRFTEELSDSLKRLITADN--ABQO
-----DMTSMVNMHTGVRGY--AAG-----DRFLEPYKAGRLQFKAQSFVDVRLTSDN--AAQO
-----STFATVOTNNMT-HI-REVQTAVEKMKLSIDMETGMRGYLL--SGE-----SFLEPFY-ANQAVFRELLOKRLITADN--PEQO
-----SARWRDHSHEVLLQLNRLTAADVDRMETGL-GYLL--SAN-----PAFLEPYEAGVKTIDETLRLKGLVADN--PDOS
NDSAKVNVNHTQIVIKKAMNIEEAAVNMETGMRGYLL--AGK-----EFLDPFY-NGTQSFNEKRLSLKGTVDN--PVOV
-----STADNLLKTESVNHSHVEVRAHMDLAAAVDME-GMRGYLL--SGK-----EDFLSPYT-NGYKTFKNEVEGDLTELVSDN--PAOV
-----DEILSAMINMETGLRGGM--NRQ-----KDFLAPYDCKGDVVARLPLSQG--I
AYTTLQOSLELRKWEHTQEVLLNLEETSSSFRETHSVLR-YIL--YRD-----XELDSFY-KNKRIVLEKIQELNKTIVN--PYBO
AYTTLQOSLELRKWEHTQEVLLNLEETSSSFRETHSVLR-YIL--YRD-----XELDSFY-KNKRIVLEKIQELNKTIVN--PYBO
GLQYQW-R-AAHFFIEHSRQVLETLDRLARVAELETBETRYL--TLD-----PAYSLEYG-VDSVSRREAAQALQMLVADD--PLOS
-----SSFAESAQWVHSYQVIALESTLAATRVESSA-GYLL--ER-----FDLHAELY-SSIPAAULSHSAQCLIT-DN--PLOH
QOMARNANQAATVWSHSQVLENAQRLKAEARNNAEASALRSHG--TDR-----PALLERMG-NGRSEAQAQVQLRFLITKDN--PAQO
QQLARLANNAAVVWSHSQVQGTQARLEAARMRTESALMARSHG--RE-----ALHERMG-NGRSEAQAQVQLRFLITKDN--PAOL
QOMAAANQAQVWSHSQVQGTQARLEAARMRTESALMARSHG--VDR-----FVLVERMG-NGRTEAMKATILAITKDN--PCOO
BROTARAEEDVRRVLLVQGGDIQTLHTQIAEGAA-VRGYLL--THQ-----EFLG-YL-NAQPLIEAALARLNTVRD--EOMR
DTAQAEEDVRRVLLVQGGDIQTLHTQIAEGAA-VRGYLL--THQ-----DFLPGY-NAQPLINAAVLRLDNIR--QOMR
BRONAAENDVROTLEVLSDLYEAHALLAETAAGVRGYLL--V-R-----DEFLTPFY-DAEPLRQVADRSLQITD--PQOA
BRONAAENDVROTLEVLSDLYEAHALLAETAAGVRGYLL--V-R-----RDEFLTPFY-DAEPLRQVADRSLQITD--PVQA
VSTEQNAQTEVQRNITVLLQQLRTSLNLAESA-VRGYLL--TQR-----DELVSTFNQSDAARRNKTREGLF-SEKTA--GKQO
VLOANKQASLNSMLNTQVQVPMDSLEDGVYRQVITAAQGLIL--AQSS-----QSEVDTHIGSFKDNAINVBPVRGVKVRLEFAG--LLFESR
VLOANKQASLNSMLNTQVQVPMDSLEDGVYRQVITAAQGLIL--AQSS-----QSEVDTHIGSFKDNAINVBPVRGVKVRLEFAG--LLFESR
AQVIGNTRVTRDRLLEQTLPAEAAEVLQSTVLNQBTGLRGYLL--AAD-----PQLEPFYVEGRQQRARVQRLREILGR--POLI
ARLLEETSVDTRDLADRLQPAQETRYRLQAGLNBQBTARGYLL--GD-----RQFLAPFYQGRABEARAARREILIGR--PRLI
-----HRTDQVSRQVADGVGFARVAAARLQALRDBQBTGLRGYLL--AAD-----RQFLAPFYDQORTEQQAADERTILVGGNAG--NTLTI
LITALEAREAVVRQVQLTPAQSAHQTMQAAYYNQDNGISYAA--TGT-----LDSLEPFYQSGRTLRASLPOLKERA--PAIT
-----ENRNNKSVTDRLALRYD-ELEDDADDLRAAILDLPHRYFLA--G--GPSRTQLE-LEGAYRLVLSI-ELER-G--VRDPDA
YGLVFNQVENLRQSYRLSRLNINVRVMQEGMDQGR-GVGRGYLL--RQE-----EF-APFYQGRQVRLSLEEARVQLGIT-RTED-CEMSQML
ENLRLIDQNNWNVQVETVSRQTLQAMKAHGYGGGMHNHKN--L-RGD-----RYKLFVVENNYYQOLSTQITAYRNL-QLSA--EEO
FLATILIMIRASTESSNHQIHADALAEITAILRQNSQ-RGYV--TGD-----STYLSYBYEGRDEYDOVSAKLSTLT--DPML
-----SMEIDVSSVITIERRETLVQGSREIKADLLQLAYIRDFIT--FAD-----QALDSLEAASADLMATLDRMIATAN--VEETR
GYFAMRDVTTTLNELATQEAELLEGSKEIQADLLRELAIVRDYIT--FTD-----DASINSLHTAENMLLVLTDEMLAITA--DSDSL
SRLATQTLLELEGIDRSYALAIQAQKLQADLLREATQIANVVL--YAD-----SRDIADYQAGTRVPEPLQOOLLEIAE--DETVR
GVYKLSDMVGTETTLTVSRAGMEKAALKEGIGILFLVLRAGYLL--AAT-----AEYDQVFAALAKNREALVKSDKEIHAAA--SEGGK
AYMKLGDMMATDSMVLRAETMEKATQEKDILILQRAERDITIL--AEA-----AAEQFAADAARIDQOALKTDEVYA-N--SEAGE
FAYVKLPOLAEI39933216-26-159--AST-----DAELOKVADEVK--DYR
-----LGGYVIVLLMVLGAGGTWGLGDFATKVVQVQIAIN--DAQTOVAAEYITLLKVDAAALADELISN--EA
STQINKIDTAYSNLLEQVRASQMASATQARARAAADITIM--TD-----PKAAEVANALAAAKKTESELMAVAAQASPAN--PKIPEFO
AGFMALVAVGLVMVAVINNLRVAVDQVAKLAFSPDSAT--R--VLEISTHLOAQIRANLVITYDAEPAMKASA--BREA
AGFSVLVVGGLMAAVAINNLRVAVDQVAKLAFSPDSAT--R--VLEISTHLOAQIRANLVITYDAEPAMKASA--BREA
AYSLKQVNDYTTITIEVEGARAALQIESVVSQOQSNHRYGVL--SGE-----REILDRFYQANQVNMRLTEETILL-S--DEEYL
AGFLVLAALMGCGVLSYSIFTSNDKDFDESEASEAYL--TGD-----AQNLRLVQSFDDTVSKVLDTASEASEAYL--DDQ
RIGLDQVWTOYELSGRQVFIQLOQRKVEYLMAEQARALNGYL--TGD-----RSYQKEFDAAAEAAALIEDIESRLQ--TAGEF
-----RDSYTSLANLSLDRDFIL--YGE-----QTHLENYQDLIKYHNQSVAEIDAKLDMLS--DNDQ
-----RDSYNSLANLSLDRDFIL--YGE-----DEYSDKNYDILSHSKSVAEINAKTORIS--ESDV
-----LRELADLSALSLSGRQLERTIL--EPD-----DALVQRFG-QGLRQADG-LAVLVRADM--SOTQR
AVSDTSRRRTVVNLRVAVASATVGNLSTVSLRGYLL--TGD-----AQGRDRAAVWADLDRTAADVDRMAEFGS--N-QNK
-----SFSAVLLIMVVMGTTVTRIAIGTGLTHSTIERKRVPAATSAH--LVNSINASLAALSGLMITGR-DKQERAAHVA--PQOAK
VEALAVNRADRMGLRMPVQGTGSATEGRMYASLAALADRYL--TGK-----DGFKTERAEITWELGAQMAEMDRVLVATN--PRNQ
FAIMGDIRAGISVSANRAYL-AGDAPQEFQGVNPMVY--MDP-----QMDANKNNAI-LTPGQASLAAILL--BDSDF
AKTFADVGRNLAAAGSQRRLYVASGADRDQKFEKPLATFRALASVGTQAILLT--DP-----EQQASAYRATARANEAFPLP--AT
-----SEENLHNFI-ERTISIRHSAMTSYANGLOKQGAQRNLIL--MDP-----ANKKGVNDS-KADEIFRLESEKILKILL--AT
TIGMCGNKDRDFRVEQDLASRAVTNLYAHGLQMGQALRNLV--MDP-----SNQAAVYNLSDSAAGFKKTSDEALVIA-A--SPADI
-----RLNHQIETFERTEHELLIQDTLKSSIAEALQIQAARN-HIN--KDTKAVENLATAMELSSHNQALN--AASF
HEMCNLADNTNKLYKHPYTVSTAALRVQGMVMMHRSMDKDAIA--KR--EDIDKAKKQVDAYEONVLDQFATLKAY--LGDI
DISTDRISALTTLKLYR-FTVTNALQAANTINIAMHRSMDKDAIA--AKS-----AEELDRAVADVDAKSVIFEFKALARERFL-G--DK
TQGLADSTASMYRHPFTVAIAIAREAKTEALVAQVMTSLVH--YAG-----AAEVSGYEQRLKARRKNDARFALLER-G--GSP
ENRNNLADLTNDLYEHFTVRKSIDRAYLNLQCMNRSNLNRMTS--PDNTMVSADMQAINDSEKFEFLKNGMIIIRER--FLGKO
GRSMLINADLANELYIHPFVATNALIQVESQINAIKRAASTMI--G--RS-SDVQRMLSLIAIHENEAAASMAVVK--RHLGD
RNVSEIGQVSEDI-TQSLKTSNAAREARVAIMKIQGIKEILLD--NPDNVYYELEKIRELDQNVLENFBIKNSNG--GNEIE
FIMGIFLHAHALSVGEENITSTRITL-NTSTIQSLOEHYIEPLNLL--EMLS-LVMS-PNESFRSTIEQELISSIAKLL--ERFL
NNEISSLOKSRNFIDH-FKVLNLTNOVEKELLIT-KARKGFTIT--NN-TYVQSLN--AEKDYKHYQDLFALLS--DNPSQO
NEISNLQKSRNFIDHDKFVLNLTNOVEKELLIT-KARKGFTIT--NDANVYQSLN--AEKDYKHYHDLFALLS--DNPSQO
NQISNLQKSRNFIDHDKFVLNLTNOVEKELLIT-KARKGFTIT--NN-NYVQSLN--SAERDYKHYHNLFSLEL--DNPSQO
NQISNLQKSRNFIDHDKFVLNLTNOVEKELLIT-KARKGFTIT--NN-NYVQSLN--SAERDYKHYHDLFSLEL--DNPSQO
NQISNLQKSRNFIDHDKFVLNLTNOVEKELLIT-KARKGFTIT--NN-NYVQSLN--SAERDYKHYHDLFSLEL--DNPSQO
NQISNLQKSRNFIDHDKFVLNLTNOVEKELLIT-KARKGFTIT--NN-NYVQSLN--SAERDYKHYHDLFSLEL--DNPSQO
QSIKSLTAQIETGQKMLPAATSMNRRLTILRLKLVSYNLAIR--DQATLDRITGLMKORYDOLKMOEKKFGL-Y--DDQO
-----KSSQSGRIHLMVQELIQTGRCLDYKQKMELEDFELK--QACEYNNVYITSLDA-FKLLA-LSA--DSQER
-----STSGIETHDITQ-NVEARIALQOLLEQNQIARDIRNELL--AS-PEASAKVQAQFQDSLRVQDTFRLQAQNPRESS--TOFER
LGLFALLVLLGGAQVQSNVAVRLVEQSSRRGEQAQIOLTAELGERTVD--ERSARQVLY-DNFVFRQRFDEHLAQSLAVIE--ALD-GHAA
IRNINIVEGHASSLLSD-FTV-ERSTQSVQASLSLRYATMLLG--EAESEESQSVQISMATDDALPTLQAITS--PQO
KVLVSQVETANSLINLDLPTVDSAESLQSVQASMSVRAHMLG--ANKNFAAEQSLSAKTIASVDVLLS--ED
FNIQIKQNTONIRKVTIERFLVVRGELVSVNQERLSLAQVLT--GN-QNARTFEEELTEASQELIEQOILIT--ED
-----QDRVARLQQRVFLVUNAVAFVSHVKDALSAQSMHTIT--E-PRFKQERQHAQOQARRQLILKLEQITV--DPRIT
-----HLOQYQGMVESTAQSFSALEAAEELSAYSVLLS-GLP-METQOR--BELAPLTNRLS-LHRQKVRRELRQQLQD-LP--ADOLS
-DG-SRIHQOLESVTTNALPLVSTTNGTSVSLAADRITFKDYLT--QD-TRMQQTYETNETQAQDAFVQAREQLAH--VAKGN
-GTSNIHQKQMSVSESMPLVLSNETSVSLAADRITFKDYLT--RTDTCQQAQREQESVAETRFKPTFQILH--DAQK
IRGNQIHTNPFESVSTALPLVLSNQTQSVQLSADKSFDFDLIT--QD-TSLMDGMRQEFQAQSKORQFQSVLAEISG--ASVNH

Figure B.3 continued

Figure B.3 continued

Ssp_52012288_26-183
 Atum_17936772_24-180
 Selo_56750264_62-198
 Selo_46129799_62-198
 Bcer_30019279_33-192
 Bant_21399048_6-166
 Mmag_23016539_32-190
 Save_29609038_91-240
 Scoe_6855387_34-184
 Krad_53768261_11-176
 Mmag_23010448_105-259
 Bfun_48780983_34-191
 Bjap_27377964_44-217
 Lint_24216205_37-185
 Wsuc_34557237_30-187
 Wsuc_34557419_38-224

Ecol_2506837_36-190
 Ecar_50120625_36-192
 Styp_16423100_36-192
 Sent_56416317_36-192
 Ecol_43218_36-192
 Ecol_16132176_36-192
 Ecol_26251236_36-192
 Sf1e_24115585_36-192
 Ecol_13364793_36-192
 Bjap_27376721_106-258
 Cace_15896003_35-189
 Msp_48832885_28-179
 Gmet_48846841_26-183
 Bbac_42522983_19-180
 Bbac_42522982_46-207
 Bbac_42522982_46-207
 lloi_56461443_33-191
 Mdeg_48864484_35-190
 Sent_62180191_48-202
 Sf1e_24113141_45-198
 Ecol_16129380_45-198
 Ecol_12515286_45-198
 Gmet_48847023_33-189
 Zmob_56426272_36-189
 Npun_33688984_25-197
 Lint_45657908_29-179
 Pres_27228663_33-198
 Dhaf_23120317_33-189
 Mmag_23013876_33-194
 Psyr_28868215_46-205
 Psyr_46188218_30-188
 Psyr_28870455_33-189
 Psyr_46187691_15-170
 Psyr_28870175_31-187
 Caur_53795565_28-186
 Rgel_47574589_32-189
 Tden_52007873_26-184
 Ecar_50122563_29-187
 Reut_53761194_30-187
 Rmet_48772113_31-187
 Ecar_50123040_34-191
 Ecar_50123255_33-185
 Ecar_50123254_32-190
 Rgel_47571710_55-210
 Rgel_47573506_34-194
 Rgel_47571655_31-195
 Reol_17428912_49-203
 Reut_46131863_31-193
 Rgel_47572127_31-189
 Bcep_46324344_26-189
 Reut_53761951_33-193
 Bcep_46311409_31-189
 Bcep_46319666_34-197
 Bmal_53723395_36-194
 Bfun_48780518_29-191
 Rmet_48770099_38-193
 Cvio_34102638_20-177
 Cvio_34105172_32-188
 Cvio_34101405_32-187
 Raqu_15077504_33-190
 RspH_46192461_26-184
 Daro_53729525_21-176
 Bpse_53722895_48-204
 Bcep_46316835_48-204
 Bcep_46323416_48-204
 Rmet_48771949_47-204
 Reut_53761244_29-187
 Paer_46164403_16-175
 Paer_15598903_29-190
 Psyr_28868700_31-188
 Psyr_23470466_31-188
 Pput_26988221_31-188
 Pflu_29611996_37-195
 Pflu_48732671_31-188
 Ecar_50120707_32-188
 Ecar_50119142_33-189
 Ecar_50119143_33-189
 Tden_52006605_45-203
 Gsam_21231332_32-189
 Xcul_39996402_33-188
 Bcep_46319877_19-182
 Rso1_17549582_31-189
 Bmal_53716753_32-193

--ASINSSAEERMAVETHLILNELGEELAIAGAE--RTDEE--RLYVMR-----GD--DHLEAFET--CAEMALEBSARDGTR--LGT-----TPEET
 --FILSARSASQERAAIQDHLDISDLAEELAIAD--RTTEE--RLYVMR-----GD--EQLHAKFY--D--DQEKHLBNTLKDIAAG-----VSPQEA
 -----CEARQVDRSLEAEVAQIIDDYQIRLLFRFNSYLES-----ROETDNKLYQGGQQLLDSIDKLKSF--VQPP-----NFBQI
 -----CEARQVDRSLEAEVAQIIDDYQIRLLFRFNSYLES-----ROETDNKLYQGGQQLLDSIDKLKSF--VQPP-----NFBQI
 --HSINKIKHDETVVDDYQYSAILLEGMLRTONSLEYNLELITASQ-----NEVANKKEIINNIEKDLKKYOSLILEYDGGFN-----LSKQEL
 --QSINKIKONTEVVVDDYQYSAILLEGMLRTONSLEYNLELITASQ-----NEGTNKEEIIDNIEKDLKKYOSLILEYD--FN-----LSKQEL
 --VLGEIK--NGPVYGRIVQVQDLVADILPPEPYIILESYLVASOALITAT-----PAELAEFKTAMTRLRKDYDDRHQYQOAGDLDYS-----VKDG
 -----CMTRSAAADVHSSQPLSAGAADIVRSIADANTASSGFLAG-----GEEPAG--TREDDIRTAADOKLVVAAANSEP-----GSPPA
 --AMNNRRAAAADDVILRSQPLSSAAADIVRSIADANTASSGFLAG-----GQETEDSRTRYDNGIRAAABGLVTA--ANSEP-----GSPPA
 --NVAVSALVVLGFLSISVQNLQTORQEEVGRAPVYITGLQQAALAAKSAATDE--GXEITGE--DEVAEESLGRCEAFDAALATARDAS-----TDAER
 --GRSARLGAALGLVLGLIGLYGRQRRRRVRENRAAQRELEAKVA-----RTADLTSTANKLAGETEEARABEGOLRSTG--ELIQ
 -----LLDAQRIAAANAGNVRMOQLERIASTELSVTVQLLDIRHAMHS-----SDDVQAAARNIDAEGRQOIRRNDDAFNSNIT--DEAGK
 --FAQK--DSQDLGNTIEYELFLSRMVSEFVDLIDRYELEVLVLVLSIDR-----ADAGT--RTMAAVRA--ADELRTVKTGAE--LHKAATGDPGYD--EDR
 -----LYRKTIIRSMIENSQKTKQNLNITLEKILACILATISSKVS-----ID--SLSEEEYKVKTLLEBETKKNVTFRE--L-----EDEL
 --NNTFQMKHHLDRLYEGSH--KIVKLQALDSLGSQD--LLVLFVFSKSG-----GV--DPEAVREIHRKLSNIOEKMGDKYGGSTH-----SPEEL
 --VIALDIS--LTFDDLE--NHTVSLTKLEEI--NNY--ANTYGLDILEKRG--ITPHQAKFAILL--EDLIDALWKSXKH--QL--QKENTRISVQKTHRRFE--QDDTQ

VATSRNIDEKRYNYIT--ALTELIDYLDY--GN-----TGAYFAQPTQGMQNAMEFAQYALSSEKLYRDIIVTDNADYRFA--
 STSRGVKENYVALNNA--LTELIDYLDY--GN-----FKKFTIEQPTORQDNFERAYVYKKAENDKLYQAGIARNDAAVDS--
 AAFLEIKRTYDIYHGA--LAELIQLIG--GK-----INEFFDQPTQSYQDAFERQYMAVMQNDRLYDIADVDNNSSYNQR--
 AAFLEIKRTYDIYHGA--LAELIQLIG--GK-----INEFFDQPTQSYQDAFERQYMAVMQNDRLYDIADVDNNSSYNQR--
 AAAAEIKRNYDIYHNA--LAELIQLIG--GK-----STSSLISRPDRIRNGFEKQYVAYMEQNDRLYDIADVDNNSSYNQR--
 AAAAEIKRNYDIYHNA--LAELIQLIG--GK-----INEFFDQPTQGYQDGFQYVAYMEQNDRLYDIADVDNNSSYNQR--
 AAAAEIKRNYDIYHNA--LAELIQLIG--GK-----INEFFDQPTQGYQDGFQYVAYMEQNDRLYDIADVDNNSSYNQR--
 AAAAEIKRNYDIYHNA--LAELIQLIG--GK-----INEFFDQPTQGYQDGFQYVAYMEQNDRLYDIADVDNNSSYNQR--
 AAAAEIKRNYDIYHNA--LAELIQLIG--GK-----INEFFDQPTQGYQDGFQYVAYMEQNDRLYDIADVDNNSSYNQR--
 AAAAEIKRNYDIYHNA--LAELIQLIG--GK-----INEFFDQPTQGYQDGFQYVAYMEQNDRLYDIADVDNNSSYNQR--
 DDLEQFEAFQQRVMDG--IDRLELVRR--GLE-----ISPPAAREWDDNQ--ANRTVRSKLNALDEALQRSYDKRAREADQJLAD--N
 QSQWTIKETSENYLGI--VDKVIIEAAG--K--ND-----SDTAANLNKS--NSQNRQKIFEEETKKVVDNNTQSRSENASNTATAANV--
 ETAERLSRAFDAPITE--IPTVDFAR--EE-----ASRR--DYFA--E--L--PRFQEKSLANALLEMQSMSEANDRABRKA--LTA
 GOYAATOENYKQYVDG--FGOLMGQIK--S--GA-----LITSQANEA--MKPVKEA--KQATEIKETSGNRAAYEESDKRTKEVDAIG--
 KAFEPKQVRAEYKEL--LMVLIEEIS--TG--G--AK-----FALAEQSIL--GL--WEIGTITMHKMSSESKMYVNOAAADD--KKADEAVTRIN--
 KAYQTAQKPLPEYAYAS--MEKVIALIRSA--B--AK-----VKEAEALIDGRYNEIGKAVRAWNGAVTBEIYAKMAKDGVIKANSSTEAYVR--
 KAYQTAQKPLPEYAYAS--MEKVIALIRSA--B--AK-----VKEAEVLIDGRYNEIGKSVRANSAVDTBEIYAKMAKDGVIKANSSTEAYVR--
 QMVDDYKQKNETVAVI--AKQVITLH--Q--NQ-----TAAEIRLSEQSGQFQAMKNVLEDEGRKEKKAALSQSDILDEIVSQN--
 SLVGDNRRNFKDKRAI--TERLSSSID--S--N-----RASAVQISMGCAATQFDEMRRVITDITLTHRKSVVLSDEIAASAAAG--
 ALDNELNARYTAYINO--LQPMKLRFA--N--GM-----FEA--INHNEQAKDLDAAYNHVLLKALERTERAKRLISEQAYQRT--
 AFDTELNRFOQAYITG--LQPMKLRFA--N--GM-----FEA--INHNEQAKDLDAAYNHVLLKALERTERAKRLISEQAYQRT--
 AFDTELNRFOQAYITG--LQPMKLRFA--N--GM-----FEA--INHNEQAKDLDAAYNHVLLKALERTERAKRLISEQAYQRT--
 AFDTELNRFOQAYITG--LQPMKLRFA--N--GM-----FEA--INHNEQAKDLDAAYNHVLLKALERTERAKRLISEQAYQRT--
 AFDTELNRFOQAYITG--LQPMKLRFA--N--GM-----FEA--INHNEQAKDLDAAYNHVLLKALERTERAKRLISEQAYQRT--
 RLLDDVITASKEYMDG--EARKLSAAD--S--GA-----D--ELTVDNAHLN--FGDVENRAKLLTQYSI--GDEHL--TAGYFNIR--
 DATRRFLANWQNYVSU--NRATLDAIQ--G--GR-----HNEGROFFW--KELPAPDQVTSALQTVQNVKREQANAKVADMERSTH--
 QKAYDKFLNKWAMKDA--HEEFLRNR--R--SLGVENLFEN--AGNANLALAAHKELEKQVEANROP--TAATVAVLEFLKMNEDLADSHVSAK--
 ETLQNFISLWTAYKSD--LAQVLSLSLE--NN-----KG--AFETIS--ISKGL--TRDSI--IKTLYSLIKSEEMGO--KEENERKY--
 RLSTELQNKLSYLDIN--SVLPTLNAMR--LD--GD-----YAGNQIL--LTLRDPAPMTYASLRTLIESEASAEELSPQADQGFER--
 QQLNAAQAARRTADAR--AQELRAILR--R--RD-----LLAGRFADTRLYPALDPLTLRMQSLSDLELIDQADVAVRADTVRSERV--
 ALLKKIEAGLAGYKNA--FDQVISMOK--MD-----FTAAVSFMTAQDSYQALMTE--TQGMVGLBARMVDAARQASADAATTRG--
 OSI--TALQDAYKAFMG--QEQIALIE--Q--NK-----LDEARMLANTMLTQGLDMQVQLRLRELNKQASAVDAAGASYAQR--
 OSI--TALQDAYKAFMG--QEQIALIE--Q--NK-----LDEARTLANTVLSLQGLDMQVQLRLRELNKQASAVDAAGASYAQR--
 ELIEGVDVTFKDYARH--AEQVHALI--A--GO-----EDAGRLLANWEMAGIAXSLAQLEGLKQLNDQASEASTGASHTYATAN--
 ALVEGLKSTYQGYIER--AEKVYTLIN--E--NQ-----AEAGRALVWGEKVMABGEMETALGKLEKINDSEAESSAATS--VYVNA--
 ALINGMTSTYQRYTER--AEQVYTLIN--E--NQ-----AEAARQLVWGEKMTIABGLEASLQKLEKINDSEAESSAATS--VYVNA--
 RQLNAVEAAWHEVVRFA--NHERFIPAV--R--LV-----SDGVSQPPYSRMNPVYATLDREMNLLVQVNC--QQAARSLDVVASSYATAR--
 ALFADISGRKRAYIDG--REAAALALIK--D--GO-----VTAVE--LLEKSLIPASKVYVLAABEDSFNEFERALVQASAAVADQGRSK--
 ALFAAIAADKRSAYLAA--RDAALKEKA--A--GN-----VLGAKKVFDEDMKPRLDAYLES--LHDLARYQKQAI--DATAGNINHRQYESGR--
 ALFDKGVGEYRQSYIKH--RDAIETKGA--A--GN-----FDRARTLFDNEFVPASNGYLA--SVLEALRDHQRAS--IDOMGKINACASRGD--
 AFAFIKILVREPYNES--RDKITKLKQ--E--GL-----TEENAVLEKEFFVPAGDAYLAEI--QKLLDIQRTI--IDATAEINRITVYNAR--
 AFAFERILVVRNPYNES--RDKITKLKQ--D--GO-----AEDATVLETEFFVPAGDAYLAEI--QKLLDIQRTI--IDATAEINRITVYNAR--
 ALLKOFODTRPPYILAA--FLKAMELVQS--D--QR-----RLOGNAI--ILNEMOPAOALFKVL--DAMMASQODTNE--IVSHAQHGQSGS--
 VQVATLEQALPAYINL--MKKATIELA--T--NQ-----HEAFRNFLITEVRAAQNVEFALQKLEKINDSEAESSAATS--VYVNA--
 ELVAELQVRPAYSSG--MANAITLAA--A--NK-----NSEAQHLLITDVRAKQDAFENALDMVYNQKELITVE--IANQSLKATNAG--
 EKLAAKMTARVAYTSA--TSEVDRILK--A--GO-----REAAQAQLIGSTLEALDAIQQRVLLMSQVQARLARETGAEEVAAIRIQA--
 RFEATLDGVESRYGPI--ALDIVGLAL--Q--GR-----RDEAVTKMNAACRPLLSALVKSANQPTDLSVVPVAKRVARADAAVARN--
 ELAKIEEVEQRYGPI--ALDIVGLAL--H--GR-----NQDAIEKYNACRPLLAELIKAA--SSYIQYASERSDAAKAA--SEVYDNR--
 ALGAEIRRIENAYGPI--ALDIVGLAL--E--GO-----REANQKINAECKPLDILVAVKVVAGVAGALSORTLQAEVRR--
 ELVAKIAEVEKYGPI--ALDIVGLAL--Q--GK-----REEAVAKMNAACRPLLAELIKAA--SSYIQYASERSDAAKAA--SEVYDNR--
 ELVAEMNRKAYGPI--ALEIVRLAA--A--GO-----HDAATARMNACRPLLAELIKAA--SSYIQYASERSDAAKAA--SEVYDNR--
 SLAAETPRIESLYRPI--ALEIVRLAA--S--NQ-----RDAATARMNACRPLLAELIKAA--SSYIQYASERSDAAKAA--SEVYDNR--
 QRIADVVRVAYGPI--ATAIVTLAA--A--KH-----AEEA--TRINTQCRLPLASLRAVDAYDLDTHRRQDMQAADYVSR--
 ELVGDIVRLEGSGYPI--ALRIVGLAQ--A--GK-----KDEATADIDNCRPLLAELIKAA--SSYIQYASERSDAAKAA--SEVYDNR--
 ELVRQIDTVEASGYPI--ARAIVDAVA--N--NR-----RQDAITMIDEQCRPLVRLVLAATDAFATYSRERAAQVLVDSGNRYTSQR--
 GLVADIDRVEAYGPI--ALAI--VNAAL--N--NR-----HDEAITMMDQCRPLLAELIKAA--SSYIQYASERSDAAKAA--SEVYDNR--
 RLVAEINRVEALYGPV--ATDIVNLAAL--T--GK-----HDEAITMMDQCRPLLAELIKAA--SSYIQYASERSDAAKAA--SEVYDNR--
 NLVAEIVDRLEARYGPI--ATAIVTLAA--E--GK-----RDAATARMNACRPLLAELIKAA--SSYIQYASERSDAAKAA--SEVYDNR--
 ALFAKIRAAQDSQRPI--FEPLYGLMR--S--HQ-----TDAARDMLENQFAPTNNAFI--SALLSLDRDRQSRDLKSMQEA--MSSQQA--
 ELLEKVKQTRAVLGGQ--YAPMYQLAR--T--SK-----GAPTYLFLKQHFASANNI--FMAALKDLAKYQ--EERMMNKAVQDSQGTYSQAR--
 GILEQALKARSALSEL--YAPLYQLIR--A--NK-----DQESKVPFLQKRFAPAI--DAFSQKFNELKAYQDGVKVEAVASNAAYSQ--
 QKIKALQDASDIFEA--KEQVALAR--A--GD-----MDGATEFVRLKLTTSQNALDILASAFANSQDQQLQAEQKKA--TADGNH--
 ALLKRLGEVMA--TGER--IDRALASSK--A--GD-----TSGAARILADPFSKRTSRAERTKL--LSDLQDRKARNIAEAAKADAFSKA--
 EILSRILGARKAAESS--LEQFESIK--A--GN-----RADTEKIFDPSFRPRMQWFD--DAVGQVQLQDRDNRDVAE--INAVKSSVQ--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 ERFNAFRAAYERYPL--LNDAVOKAR--T--GA-----PDALAAAYARVTPAWEVIRHANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAKTLPAWTEVVENANV--LVOENRRFADQOSATL--RESVHGT--
 QRFVAFRGAYDRYVPI--LNEAVOKSR--S--S-----REEAAVAYAK

Psyr_23469129_31-192	QLLRGIKEIEQOTLSS	-TKSVIALRR	-A	-GD	LAGAQAALLSQTSGNYSEWLKRNALIDHEEASIRVQLDNVQATASQFR
Psyr_28870840_31-192	QLLRSIKIDIEQOTLAS	-QORVIALRR	-A	-GD	IAGAQAALLRQVSGDYSEWLRRVVALIDHEEASIRSLQDDVQATASQFR
Hhep_32262705_40-203	SLILRSIEETQAKAIP	-IVQIIOAKL	-A	-GD	SQKARTILITLSPYFTQWLAEINEFDYQENANSGLTHQLRSVDSFEFR
Rleg_4973017_32-195	TILSEIADIOAKANPI	-VAQIIALQE	-K	-GD	GEAARKILIEQARPAFVAVLGAINKPFIQYQALNKSIGGEVRSSASGFK
Rrub_48763002_15-176	TILGOIKQTESRTMPI	-VKTVIDRRR	-A	-GD	LEGARALLMAEARPAFVFWLERINRFIDQENANVOIAERTRAVARGFA
Xaxo_21107857_11-168	DILQIKITIEQRTPL	-IAQVRFAFR	-A	-AD	KAAQOQHLLQARPAFIWLASINAFIDLOEAKNRQAARQAVATARG
Xcam_21231495_11-168	DILQSIKAEIQRSMP	-IAQVRSRLR	-A	-GD	RLHAEEDLLKQARPAFIWLASINAFIDLOEAKNRQAEAEAVATARG
Wsuc_34557253_30-190	TILQKINEIEKRTPL	-VQELIRLKL	-Q	-GN	QEHQAOTLLMKEARPAFVFWLVINEFIDYQESLNQOLTPIARAEAGG
Vvul_27361676_33-193	AILQIRIGDIQTKTLP	-VREVIAXKK	-S	-G	EDVTMMLIEQVRPSFANLKVINEFIDYQESLNQOLTPIARAEANGFO
Vfis_59713711_33-193	SILAEINRAKNKTLPI	-TENIIDDKK	-S	-G	SVMGAILDSARPFIQWLTIINEFIDYQEQYNQVLTFFARYIAGGFO
Vfis_59714255_9-170	SILANIEKIQTKTLP	-IDQVISDRK	-S	-G	KLITDVLDSIRPAFIQWLQDINQFIDYQETIQNOLITSEARGTAGGFO
Vfis_59711698_9-170	SILANIEKIQTKTLP	-IDQVIRDRK	-S	-G	KLITDVLDSIRPAFIQWLQDINQFIDYQETIRNQILTPEARNVAGGFO
Rgel_47574745_31-189	AKLAKLMAARADYRGR	-RDKVLELLR	-A	-GQ	MDPAKLVLRVETPKQVAYNARLDELILQENLMTASADEVSAVSSIT
Gsul_39996396_31-188	AKLKAVEASRAAYRED	-LKLVEYLIR	-A	-GN	KSAAQKMLFGSYRERQRSYFDVSLDRLFOVQAKLLAVSGKEAQVTVSSR
Gmet_48845935_32-188	GKLKTVTDELKVYREA	-LATVIADIR	-A	-GK	YAAQAVLLITTVRDRQATYMKGVDELVQVHRAAEVKLGQGVYAGKRAQ
Daro_41725108_32-188	ATLAKATEARKAYLEG	-DEAALELLR	-A	-SK	KDEAVDILLTKVRKSOADYIAAVNELIAPQSAMMEKSGKNAEVMNNAR
Tden_52008473_33-190	RLMAEL	-FANQVMEAVVEALL	-D	-EN	NYGALTQLOREPLONRLVEALDNMTNLQRKAAVEALGKFDAYQATK
Tden_52006559_34-186	SVYDKIAAARATEPFA	-LDKVRFAFR	-A	-GN	PLEATRVLVEEVRLQOONLAALEEMAAQLEGGAAVAMVDAAADATAN
Xory_58582465_75-232	ALRANIDKHRNEVKPI	-NTKVIDLAN	-A	-GQ	SDEALSIMMTKSAPAMQRQODATQNTITQDKSAAADAAPQASMSDESR
Xaxo_21108106_1-131	ETRAIEDRRREIVKGL	-NDKVITELVN	-T	-GH	SDDALPLLLTKAAPAMQOQDKTAEINLQNLKLAGDSAAALQMSMODESR
Xory_58582471_51-208	STRABVDSLTPKPVRE	-NNKVIDLVT	-T	-GH	SQDALPLLLTKAAPALQWQDKTAEINLQDLKLAEEAAVEALEMSDDSR
Xory_58582468_27-183	QMRQTIDIAREKARIA	-NQQVMDLGL	-N	-DK	PDEALKVLMMQQAAPANQWQQAALDAYAARQRTLTGTAACDDANTAMDHGR
Xaxo_21108103_1-133	EMRQIDIDAREKARIA	-NQQVMDLGL	-N	-YK	PDEALKVLMMQQAAPANQWQQAALDAYAARQRTLTGTAACDDANTAMDHGR
Xcam_21231323_1-138	ARRDKIDASAASAAKRA	-HAQVAELGL	-A	-SK	SDEALAMLMMQQAAPATQWQQAALDAYAARQRTLTGTAACDDANTAMDHGR
Xory_58582470_36-191	ATRAKIDASNKAARAL	-NQQVIDLGL	-S	-GR	TNEAMPPLLQRAAPATQWQQAALDAYAARQRTLTGTAACDDANTAMDHGR
Xaxo_21108101_1-135	ALREKIDANNKAARAL	-NQQVIDLGL	-A	-GK	ADQALPLLLQRAAPATQWQQAALDAYAARQRTLTGTAACDDANTAMDHGR
Xcam_21231752_31-190	TKLAQFDRAWQKYLDE	-RGRFTEAAN	-E	-EA	LEANPELAELSRAVRASSEDVTLMTDLSGLRKSASASANAETGAHSSS
Xaxo_21108705_31-190	NKLAQFDRAWQKYLDE	-RGRFTEAAN	-E	-EA	LEANPELAELSRAVRASSEDVTLMTDLSGLRKSASASANAETGAHSSS
Rmet_48762922_34-192	ALFKQYETLLPPFRER	-MKQYVELIGK	-Q	-EN	LDTSQFESIVFTESAELIKDSHALEDLLIKIKRRDERAKAMNDGETSVYN
Reut_53762135_34-192	ALVKQYQOTLAPAFER	-MKQYVELIGK	-Q	-EN	LDTSQFESSVFSSEADLLKDSHALEIMATMVTNRDRDRARSNMEEAQVHFR
Rrub_48764797_28-185	EKNEAFRVWNGEFALE	-DEKVRGFRV	-A	-GE	PLKAQELSVSEGRQIVGDEKONMDLVLDINTQMEQAOALDAYESSR
Rrub_48764795_29-185	NAISAFPTLWQYATIT	-OSQITIALAK	-S	-GR	MEEAIALSRVDAVARVLDRADTTLTALTESERAAMTRTEAQASQVQAEAR
Bpse_53721497_32-189	ALLAADRARVSOLDAN	-RENVLALS	-R	-GR	KQEAELMGTMTETLAQONTAALAAHRAFNVDLGAQGSNEAKDIIDRA
Bcep_46310707_33-189	ALVDELITTRYTTIVKE	-GVPEFEFAAR	-A	-GD	MAAYHVAADTKISPMFVAYDQAAASAVIASLOKRAEDDQAATQSGIT
Bcep_46322311_35-188	ALFDTLQRRRTTIVKE	-VFLKAMSOLD	-D	-DN	AFDFLETHRTAPPGLFTAYQQAIDALESFQKRBADDAAGARFHE
Bcep_46315673_18-170	ALFDTLQNRNRALLDG	-VEVKALSOLD	-D	-DN	GGFGLDTQRTAPPALFVAYQQAIDVLESFQVAREKFAEAGVHFR
Bfun_48786542_30-189	RLIKEMQRRDAFLHSE	-AVEPALFAFR	-D	-ND	RAAFQQLQAKHLPSLYSAYEKAMLALEQLQLDHGAQRYQAQDL
Xaxo_21108114_1-133	ELYDEVKLLSDYKIA	-NAALSAARV	-A	-GD	LVTANRVSDQSRPARRDLFAKLEVLTKFNVAHMDSEIAQAQATYRGR
Bfun_48787402_62-215	FLLEAAGKRRDITMKR	-WRAFIDALK	-S	-GD	DDAAQKIGNSRLSPLFNDMSDTNDRKLSALYANAKKRSYNEARVSS
Bfun_48788521_31-187	RLTDDDLDAKRTAVRDR	-ETDKLIALAR	-S	-GD	ASWMDSESRANHLGLYAMNASSGALENVLDQQASDANERSALFHT
Rsol_17548524_29-187	KVAGEAKKQORDLIKE	-NIAAAVEALK	-R	-GD	KDAARQLAVEQIPINFRTYASQSDRLNAIQTESAALYDASQSFYFKG
Bcep_46321947_17-173	RAADANTKFNALVER	-SLEPMFTALR	-A	-HD	PANTPNVRHPIPPSLFDIASDSMDALGRIGQMSASRVTTDAQARFHT
Bfun_48788500_33-185	RLAQDVVSRRREALHQ	-LDAFAATIA	-A	-ND	QAKLVDMGKELQVAYNDLANADALRKYQFTSAKEGYDAASSSFT
Bpse_53719442_41-196	RLAQDLASKRQILORE	-LDAFAATIA	-A	-ND	RDRILESAKRMQNVFNDPSILASALRAAFQKQSVNFSQADQSVYASR
Bcep_46320035_33-188	RTAQDVAAKQKQALQRE	-CDAFGALVT	-A	-GD	RDRILEGAKQLQVYNDLTASALRNFPQSDAQRGYDHABSYETTLR
Bcep_46312035_33-188	RLAQDVAAKQKQALQRE	-CDAFASVYG	-A	-NQ	ADRIGEGAKQLQARYNDLATASALRNFPQSDAQRGYDHABSYETTLR
Rsol_17549625_31-187	KLADEADALRTVFLTR	-SAQALMQATIA	-G	-GD	AERTSQTYMEAMPALRYPLGDKVATLSRMQMETAGSYEAGQRHGG
Reut_53762140_18-175	ALADELARKRADFMTR	-ETEPLOKNSA	-D	-GN	QEEAVRLASKVLPPELRAMSAAHLEKQKFLTKAFANQSDQYARFETR
Rmet_48769257_19-176	ALAGELAAKROAQTFF	-SVKPLEQAMV	-A	-GN	HDEALRLARDVLPDLQGRMSAAHEKLEKFPQDTGKAFNFGQSYRTETR
Rsol_17428475_31-188	RLSREVGTKREAAAGS	-LRDITKANFR	-A	-SD	RAAADIMDKRVSFKSFREANDASQALGKQQLTFKAFNDDSQYARFETR
Reut_53761466_20-177	NLTQDVSAKREALFSC	-GVAPMVAALR	-A	-GD	REAVMTSMVKMTIPKLDIALTAASGDLSSRAQIATSAHRVYVESQQRVY
Rsol_17548728_31-187	RLAADVDAKRTTFLDR	-ETEPLOKNSA	-D	-RD	AAAVDKAVMTVPPMFVALSAVADALDRNQAEQAKAAYEGVATRSQN
Bfun_48787377_20-177	VLADRVNMAARTALLQI	-GVQPSIEALAK	-GE	-GR	HERADAVMKIMSPSLSLATNSADADLTQWQKARGQAFADAGLHBR
Bcep_46322449_38-195	PLASRLDAARQALIQG	-ALKPMIDAMK	-G	-GR	RDDADRLMTVAPPLSVLAQAQATDALDAYQAARGDVYDTAQTYNNWR
Bmal_53716899_85-242	RLAQAQANAARALADE	-ALKPMIDALK	-GR	-GR	HDEADRLMTVAPPLSVLAQAQATDALDAYQAARGDVYDTAQTYNNWR
Sone_24376324_30-192	QLFKILGEAAEKYFSA	-HSSLVLAAL	-Q	-GD	MASANIMIKITLTROTLEVAGEETMNLRHENDRAQAEQVATSNAYKTAK
Cvio_34104168_26-184	KLQITQDREDMKLYRQ	-RDKYLOLQI	-T	-NN	VAEADKLMTGDMSEAKALNQALTDHIFNYKLADOLSLKDNAAAYTAA
Psyr_28870444_26-185	NMLAADRAAVKHYHRD	-IPAFPERSR	-T	-GD	YAGAKQMITAGELFKASLARLRTVAHELYNTQGGIATVENNKQAHRAS
Psyr_53693339_33-189	QAFDELNKAISDYOTA	-QNHYLASVA	-A	-GN	FEEAVTISNGEMKNAADQVENTLKKLIGINDGKAERAGNOADAYQOT
Pflu_48729692_32-188	QAFDELNKAISDYOTA	-QNHYLASVA	-A	-GN	LEAAVATSNNGEMKNAADQVENTLKKLIGINDGKAERAGNOADAYQOT
Psyr_28872674_33-189	AMLEGLSADTAKYLSI	-LDQIKQIDV	-A	-GQ	NDQAYARLTNELAPQGTVLDTKLEQMITLNOQGAOTAAKSAAMQYQA
Psyr_46189178_11-168	SQFDDARNKMSNYLGS	-LEQVIALMD	-A	-AD	HEQAVSEANSEQADRASAYQETLTITRDENAKAEQSGADATSVYNHVS
Psyr_28871675_33-189	QAFDELNKAISDYOTA	-QNHYLASVA	-A	-GN	HEAVSLANGEQALKAAYQETLTALRGHNAEAVVSGKDATAVYDHS
Rsol_17549061_29-185	QAFDELNKAISDYOTA	-QNHYLASVA	-A	-GN	HEKALSPANAEQOVERANAYQETLTITRDENAKAEQSGADATSVYNHVS
Reut_53760762_34-191	RLKKLQDTQERYQAL	-RAKVIATIN	-A	-GD	KEAARDLITTEMPAVQSTIETDALDAVAYQITQMLVDTTIGRALTASER
Rmet_48770042_26-187	RLYKDLAAADTRYNSR	-VAIVIDGVR	-A	-GD	AEKARAEINGLSAAQASYFAPLDALMEVGKAVYSAKESAEANEAYRSK
Rmet_46131652_18-176	AVFEQLTQARATYDGO	-LDLVMRQLG	-A	-GE	YDGRAGLTATLPSQAAVFDKLDALIAQLGQKLAVEAVEDATARYAETR
Mmag_46203065_28-185	KTFSDILTSAREAYNG	-DMVLRQLG	-A	-GE	FAGARAAVLVILPAARQPYFEKLEDMGGGNRLALAAVQADAGFYVNR
RspH_8250660_26-184	RAYDALRNKLGSLADQ	-WRLRAFHE	-A	-ER	YEEAMAFRGPMQANYLSASTAARGLIDINLAASRTADSEIRRAQTDR
RspH_22958341_26-184	ALLDEYATLRKIGISEI	-NNKATIEFSK	-K	-ND	LEGASNLDDPDLATQKRRBELAAAVIAQQLBALDAERDQVQAIMDEA
Cvio_34103819_33-189	ALLDEYATLRKIGISEI	-NNKATIEFSK	-K	-ND	LEGASNLDDPDLATQKRRBELAAAVIAQQLBALDAERDQVQAIMDEA
RspH_7532754_28-185	ARLAEVQEAERLARI	-DEKATEMAR	-M	-GD	HAQAVTYLLKEFAPANNLITLALDEMGAFQAOQMDLSHEAAASAEISR
Pflu_48730667_20-182	AGDRMEKAWPYQAI	-YQYILALMK	-A	-GD	GYEGFTIVVTQGREQWLAMETRLALLAHTQQQLTADSAEAQRQGISR
Psyr_46188223_34-194	AGDQMDRDPWGYQAI	-TEQAKSVAIL	-S	-GD	LENGRELLDGLQKNYRVVMDLITMTNSDMQVSEAAKRSVLTESASQ
Psyr_46187354_20-176	AAGDQYDALMPAYMSA	-SONIFDLQO	-A	-GK	IAGARAIIDGDRSGYGLKVMQDLTITVNSNNRQAGEAAVASDQTSNSAQ
Psyr_28870740_34-193	VAGDQMEQMLPAYIAG	-SQQVVELMR	-A	-GD	LEDARNRLNALAEGEFNKARGYLQIMIDSNKRIKEGAEAADRLQSTS
Psyr_53693337_34-193	AAGDQLEMLPAYISG	-SQQVVELMR	-A	-GN	YDNARTRLNALADGFPVKIRGYMRTMIDSNNRQKEGAAIAEKLQSS
Gmet_48844820_32-197	KLAAEFETAREAFVQE	-GLIPVREAVIT	-T	-GN	YEAARTQLNSLSSESFTKVRSMYRMTIDSNNRQKEGAAIAEKLQSS
Daro_53729524_186-351	QLADAYVAARKVYVTE	-GLLAAKAIAI	-T	-GR	YKDAVELTLKKVNPFLKPAEALERKMIKEFAAAREDFERDDKTRTDR
Bcep_46321623_26-184	GLVAQARDGAANYFAA	-IDDTVKMKAD	-G	-GK	FDEANDILLKLNPAAYEASKRADDLVQLQISRGKTQLEETDKAYQQR
Bfun_48782649_13-171	GLVQANDSLQNYFAA	-IAETAKMKKA	-E	-GR	AEMAQAYLFANVAQYRDELEGIVETLRVEKNRQKDDAISALNGLMATA
Reut_53761418_15-173	GLVEQARESGFNAYEA	-IRETAEMKKA	-A	-GR	NELAQAYLFANVAQYRDELEGIVETLRVEKNRQKDDAISALNGLMATA
Gsul_39995862_33-195	BOIAEFKNGYDAYVAG	-GAKLAEMARSAASGD	-AS	-GR	DELAQAFLFANVARYRDEMEKIVOTLRVEKNRQKDDAISALNGLMATA
Ecar_50119387_32-189	ALLKIVIGENSGFGNS	-NATLIDTIV	-R	-GD	REEAVRYAVTIAAPLYNKPAQALAEVLEISKEGGEVYDADNASYRRS
Rrub_48764496_10-167	EAYKTFADNYAGFAAG	-YIRIALDSLK	-A	-GR	RAEAVAFATGSVAPLYDNPAKIASLVANVQDKSAMYEDMASYRRAL
Rrub_48764497_1-152	EAYGVTDFNAGFASG	-YIRIALDSLK	-A	-GR	LAEASKISGASSSKYSQMLKOLAKVEMELATAKNIVASADSYRTSQ
Mmag_46200981_38-194	AKLMRFKQAFQYRDN	-TERITLQK	-S	-GQ	IDAEARELTGWTETPPYLAIVKSTDRLETSQIOQAASAAINDHTNRIR
Cvio_34102727_35-196	DGLKRFDAWAAMEGH	-CKQYRDLSEYDQGN	-EN	-GD	IDAEARELTGWTETPPYLAIVKSTDRLETSQIOQAASAAINDHTNRIR
Sone_24372094_36-190	RLADFDROWPPYKAK	-AKKMDASY	-R	-ND	ADALAFISAEQKVGREANEAALQIVVAKERKSAEALVASHKQARQ
Naro_48849030_29-187	GAVATIKREWARLLQG	-QYTHQAMAL	-S	-GD	NKKALAIMSKQCRPLQQAADDVTELSNINIKLADKSLDAGCKTQARQ
Wsuc_34558217_26-182	QIYEKEKALYAVYIA	-ADEVLRLSK	-E	-GK	NEAAMQMKSEVTPFQQAADATLBAIVNNKVDLGGKAYDDSPVYALQG
Cace_15893832_34-189	DYFNEVAELWEDYSKV	-YOKSMELSR	-E	-NK	QAGAQMHNSGLDSFYAVEDALAEIVNQKAAADVASQSGSEETASAR
Ecar_50120989_26-182	LEBQFQSMLYKQYND	-YITLAINATK	-A	-GD	KLEGKEYMLTNRLMSRLTDAIDHITYNENLAEKNAILANLKKEA
Psyr_34557239_33-189	IKVKECAESFKYYSOI	-VEKVLLESR	-K	-GM	NEEALKLYWERNMGDSNRKQKLDMLKTNIDBASQQRQYTKTIVTSRGE
Rpal_39934745_190-343	AVADAFSAARQAYFDE	-VVGPGVGMVAA	-RE	-GD	KDALENIINHILKSEYAPLQKADRLIKLQSDVGGELIYSESSQSFNN
Tden_52007871_1-135	ALADAFSASRQKAYFDE	-GVVYVLAALS	-H	-GD	NDQATKISONEAKIMHDCREELMQLVASENENLSIDKRETNILEYSGM
Sone_50261353_32-187	QLTQQAAPLFTAAQNR	-TEQLVSYLRSFN	-A	-GR	FDRLGELMTGRABDLFRAAKTELDKLVALVQQAARDEYSSGQRYT
Mmag_23013720_30-169	ALARDABAGIKGYQAR	-QESWDRIA	-A	-GN	QQAQETLQGMRETYPAVRANAEALIASQLQAGDESPFAAQSRFVMVR
Mmag_23013426_39-195	ALARDABAGFKSYRVY	-ETANWDRIS	-A	-GD	VAHQINDKILPLXQVDPISSTISBELTALQITADKTEIADVDELVSS
					ANAVILVDMKARGVAKFREAAATPLKRLDT
					LEGARIDLLGPGDQDFRKAATPLKRLMDYQREANASFLLEGANYASDR

Figure B.3 continued

V	ADRLSIBONKVVQAA	GN	GN	LNHAA	N	GN
V	VLADRLSIAAAYREI	AGRLTALK	A	GN	GN	GN
V	RLADQEAAMNANRS	VOKRLQIVQ	S	GN	GN	GN
V	KLHEAKPLIKVADAT	ERLBRASLT	S	ED	ED	ED
V	RSVDALAKLVEAQDF	ERLKQILE	R	ED	ED	ED
V	RSVDALAKLLEAQDF	ERLKQILE	R	ED	ED	ED
V	KLEPDAVLSWDDMKAT	AGPIROLAN	Q	AQ	AQ	AQ
V	DLYRQESQWAXRAYL	NRVVLVLSQ	R	ED	ED	ED
V	KLFDFVADMKKASAS	ASVABEMIE	L	NL	NL	NL
V	KLFDFVADMKKASAS	ASVABEMIE	L	NL	NL	NL
V	AWDELTKSVEYDEYTL	STKLMSLEL	D	GM	GM	GM
V	AW DELT FVWORJQAF	LEQVLSASA	A	CG	CG	CG
V	QAYDAFSALDYOQBD	SERIVRLMK	A	CG	CG	CG
V	QLYDQFKRTFAAYRTQ	TAQSTFLAQ	Q	CK	CK	CK
V	TLNDQFLVQVQVQRTA	LDRSFVLAE	Q	CK	CK	CK
V	SLTDQFVQVQVQVQRA	LDRSFVLAE	Q	CK	CK	CK
V	ALDFRQFMAHVKLFEE	QAQVMTLSA	Q	CK	CK	CK
V	ALYQFSEFTLNDNVQA	QONQMLESR	Q	CK	CK	CK
V	AAPDQVQGLNLQVROL	ESMRMTLSQ	A	DR	DR	DR
V	NLYQVQVQGLLQVROL	ERBLKTLTR	A	NK	NK	NK
V	AAVQDQVQGLLQVQRT	EDRMKTLTR	N	NK	NK	NK
V	ELFNIVQVQWQVQVQL	SEKITITLSK	Q	LR	LR	LR
V	YLFYTEREFTWYAGHA	TVAVLEHSR	E	CK	CK	CK
V	NFNDAWKVMGMDYDST	TARTLIDLSR	K	NE	NE	NE
V	LIFENNFKLQVKEYLST	QOKLIDLSR	E	NK	NK	NK
V	ALLDQFVQROQVEYLEG	NAALLALSR	D	NR	NR	NR
V	ALLDQFVQROQVEYLEG	NAALLALSR	D	NR	NR	NR
V	SIVERYLTAOTROLLOS	QOTLVOMSR	S	KP	KP	KP
V	GNFEKFLDHTQVQYRA	MDRGMLESR	V	NK	NK	NK
V	QMFSAVASASWASYLVK	STNLFELSR	Q	CG	CG	CG
V	QMFSAVASWASYLVKLV	SHDLDFQSR	Q	NL	NL	NL
V	QMLQFSTKQWCEYLAL	EPKVLELSR	Q	NK	NK	NK
V	ALYDRAFTWQVOTYLAL	TEQVIELSR	T	NK	NK	NK
V	SLYGVQTKAWAAYLVNG	QVDVVALSK	V	NR	NR	NR
V	ALYETWSKLWDDYKRS	ADBEVFAVRSK	VR	CK	CK	CK
V	QYQKVKVBAWGVKYL	VPKLIDVMSK	VR	CK	CK	CK
V	FLYEEKFLNANVQYRL	AKKIVEMSRKSV	CB	CK	CK	CK
V	KYDYSVLSWNTNYDLS	APTGLEMSRKA	CB	CK	CK	CK
V	EDFLAVKSYDNYDLS	LODLEFLA	A	CD	CD	CD
V	SVQVTFYASQAFSAI	LDVLELQI	K	CA	CA	CA
V	KLSEYEBELCKWYKL	NEQFNKLVN	A	GL	GL	GL
V	ELKPVDAWLAWEKDL	CKGVQVQSR	T	CK	CK	CK
V	ASVKEIAGWHEKFSK	CEGLTMSANY	Q	CK	CK	CK
V	VLATEICHRITLESIGI	VRSSELRIS	A	CD	CD	CD
V	SVATKINDTLTPENDNA	ANNAVADLR	A	GN	GN	GN
V	UMADQLOLTLAPADLI	KVAFHALLI	E	GN	GN	GN
V	AYFEKLAADMKSSASS	VISDTSSALDD	AE	CK	CK	CK
V	AYFEKLAADMKSSASS	VISDTSSALDD	AE	CK	CK	CK
V	AYFEKLAADMKSSASS	VISDTSSALDD	AE	CK	CK	CK
V	BLJGEFKPREEAVYRDL	REKALTAVA	L	ES	ES	ES
V	KLEFQLOLTQVNTYMDI	HAQIESGR	T	ND	ND	ND
V	GIYNRLRKSSSDYRST	HNBNINAV	K	ND	ND	ND
V	KIFTOQVLEBNLGNNA	KDRFPELST	A	GN	GN	GN
V	BLJGEFKPREEAVYRDL	REKALTAVA	L	ES	ES	ES
V	KLSEYEBELCKWYKL	NEQFNKLVN	A	GL	GL	GL
V	ELKPVDAWLAWEKDL	CKGVQVQSR	T	CK	CK	CK
V	ASVKEIAGWHEKFSK	CEGLTMSANY	Q	CK	CK	CK
V	VLATEICHRITLESIGI	VRSSELRIS	A	CD	CD	CD
V	SVATKINDTLTPENDNA	ANNAVADLR	A	GN	GN	GN
V	UMADQLOLTLAPADLI	KVAFHALLI	E	GN	GN	GN
V	AYFEKLAADMKSSASS	VISDTSSALDD	AE	CK	CK	CK
V	AYFEKLAADMKSSASS	VISDTSSALDD	AE	CK	CK	CK
V	AYFEKLAADMKSSASS	VISDTSSALDD	AE	CK	CK	CK
V	BLJGEFKPREEAVYRDL	REKALTAVA	L	ES	ES	ES
V	KLEFQLOLTQVNTYMDI	HAQIESGR	T	ND	ND	ND
V	GIYNRLRKSSSDYRST	HNBNINAV	K	ND	ND	ND
V	KIFTOQVLEBNLGNNA	KDRFPELST	A	GN	GN	GN
V	BLJGEFKPREEAVYRDL	REKALTAVA	L	ES	ES	ES
V	KLSEYEBELCKWYKL	NEQFNKLVN	A	GL	GL	GL
V	ELKPVDAWLAWEKDL	CKGVQVQSR	T	CK	CK	CK
V	ASVKEIAGWHEKFSK	CEGLTMSANY	Q	CK	CK	CK
V	VLATEICHRITLESIGI	VRSSELRIS	A	CD	CD	CD
V	SVATKIND					

Figure B.3 continued

PLVQSKAEKSNATYNA	LDITVTKM	D	GK
PLTRLDAAARQALIG	QKPMIDAVE		GR
SLDDLDGARYATLAK	VEVEFEFAAAR		GD
ALDDELAARYATLAK	VEVEFEFAAAR		GD
ALDDELAARYATLAK	VEVEFEFAAAR		GD
ESASUNDEKRYQYQA	LAELIQLQ		CN
NIREDIDRRRAETKV	NDKIVLVEL		GH
ALDDVLRKRTTWRE	VEFEFAALN		ND
ALFDALQDRQDELIS	VFNKALQQLQ		DD
VLLGQIRNTAMETSR	QDFPMFQL		N
QATEKLQCGQILGYA	LSLQILQLG		GG
SGAKKLKQSYBEFSA	LITLFINME		N
ESGLVFFRGFGESYVE	STRALILKQATARG	VG	
SSNLNRKLKADVDVY	LHELIVLEH		
QOPATFRKRRIQFVE	RHELVRGV		IN
DSILDRLTQARPAYRD	NKAVELVQ		S
MYEDISIKSNATTKK	YENIINOVLK		GB
SMIEDIAKTKNNVQ	YKRIIDTVWK		ND
KLLAAIEDQVQVRLA	ADERTELLS		S
ALDEDILIKAYTNRYD	LIVLMOKATS		GE
ALDGDRLKSYDAYNG	GIMMLTAAK		Q
ALVDPLKKAYQYQRI	GKPMFLAATK		GH
KILDIHETRAAQSSI	YKGYFEVD		R
VEVAARANAMNTWRAS	HESLTNFIN		S
ARNDELKARYSDYVE	QTKPMFDIA		S
AKVAFEFGRGYALWSA	SARVLSLAA		S
AKVAFEFGRGYALWSA	SARVLSLAA		S
VLPKDAQDQKFAALIA	MLMLKNLK		S
AALFPADQADQAFAKI	RAELVRLEK		E
GNFEVAQKAAEDFTRI	TRTELVELSR		D
FAPFDILVERAEFEKTI	RSETALRGTQ		AP
ATPDNLSSIRKFEKFTI	RIETALARARA	B	
KALTVOEKRDVSVKI	NMLLDYGI		R
KIINTEIRARQOYLES	RFRILKQD		S
KIITEIRARQOYLES	RFRILKQD		S
KIITEIRARQOYLES	RFRILQDIO		S
KIITEIRARQOYLES	RFRILQDIO		S
ESQFQIEKNKHRYGL	LEKALIAVE		S
AISSMDAENKNALMQ	LIDPEFAALN		CN
EASERLTPLLAEADN	AKLGRGLQKK		D
AIIRDITLPEDASNA	TEAKRGMAKI		
DRLEAARAALDHWRAL	HORNINSLA		CN
FDLQAGDQLLVRWKQA	NDNRINSYIS		V
FDLQAGDQLLVRWKQA	NDNRINSDES		V
FDLQAGDQLLVRWKQA	NDNRINSYIS		V
EDLGQVQDQLLVRWKQA	NDNRINSYIS		V
NALASYEDLWTSQIR	NKDAILLSE		E
NALASYEDLWTSQIR	NKDAILLSE		E
ARVTALEARYALRAGA	LAVLQVQID		
ARVTALEARYALRAGA	LAVLQVQID	A	GN
ERAAALDRYRYKYHA	LVDLVQVLE		
ERAAALDRYRYKYHA	LVDLVQVLE		
NIQTLEKRLDQYQV	GFEPALQQLK		EK
RIARAQAQAATFVBE	VLEPEKALV		ND
KLQAQVEQARSALVR	VEFEFEKALD		FD
ABYKRLADASTRYFAP	NKLKRLVE		Q
ABYKRLDASTRYFAP	NKLKRLVE		Q
AKPDAAKEKADAFSKV	RSELQLQRLK		D
QLDQATKNSYQVTRIA	LELIDFLI		E
NSSWMPMVRATIRYA	LELIVFLIE		N
VLASVSTSYKEYNIA	LELSLAPL		E
ALQKETKESFARWIND	LEHQATWLE		
GLOKETKESFARWIND	LEHQATWLE		
GLOKETKESFARWIND	LEHQATWLE		
NIGLEQLOKTEFATY	GRQVALSI		N
AKIKAFGAVALWSAII	LRDFVKLRQEPDLIS		
ARLRAFSTAVTWSAII	LRDFVELVR		Q
ARLHFAASAVTWSGII	LRDFVTLVQA		Q
DOOKFEAAVATWSGII	LRDFVTLVQA		
ALQKERATAGELMINI	VALILGQPLDLOQEP		
BITESLVTPLVRKVR	TLAYGEAARA		GS
BITESLVTPLVQKVR	TLAYGEAARA		GS
QALQADAAARAFVKY	LELPAITALB		SD
QKALIKGALLSGYADQ	EKGVSAQVEAR		
VEVAGLKLGLDYANG	KPKVHKHIA		EE
KTYEEFLACWSDYQA	SEKFEKAAQOE		
VEVAPARKLADYQAQ	FEHLSIALE		K
WEMKPKKIKDSFYKI	TEQIVLILLO		Q
AFRAMKPKIKDSFYKI	TEQIVREALL		R
QFVDTANGDMTEWKK	HSIARITQE		N
QFVTAARNTGMEQKSR	HRARARAD		S
KIYASATALEGAYAR	LDVAVRFLP		GE
ALRIOTIRATNEYVAG	FSSQVAVRMD		HN
RLDVLQAHADYVRA	FESELAAS	MD	PD
QJLDADDALESEYAAQ	FQVLVQVIM	GD	VEFMD
QJLBSLDDLEQOYFI	QKTFPDPLKA		Q
ERNRSTAAVE LAP	NATRLDSD		S
EATNEMETIGNDYDR	AGVYPLVQ		A
FLVNDHITARTNWISE	INTIIDLK		K
ALDQAQSEAFQOMYDV	LQSTHYIK		
ALDQAQSEAFQOMYDV	LQSTHYIK		
ALSKKLETSYATAYDK	INDIVDSIS		S
KLDNIQAALNEFGVIR	YKMLNDITQGLISRR		
KENAMNLTFSYDMSJ	RDKMTNRVAG		
AALARIENIRIETAVTR	RNLDNEHAKLQV		
EVFEKPNQIKKLYRSIT	RBDINLNL		NN
ELKROLVASSDGLAAL	LARADALIR		N
ALQRAHQAQAQAHAAD	VORAMARAL		AD
SARDGMAVAVARAHND	VORGEALIR		K
SKSKEVENAFELVGO	GLRFLAALAE		GD

151

Bpar_33596128_48-206
Bbro_33601527_48-206
Bbro_33601934_48-207
Bper_33592680_32-191
Bper_33592488_30-188
Ecar_50122545_31-191
Ecar_50122513_30-192
Bbro_33601551_22-186
Cace_15896714_38-194
Bbac_42522860_32-187
Bbac_38327424_31-148
Bbac_42522674_31-188
Gmet_48844408_26-190
Gaul_39996049_29-191
Ypes_22127066_57-216
Ecar_50121245_89-248
Mfla_45520525_328-462
Llo1_56461335_28-193
Dpsy_50876785_41-125
Dpsy_50876786_72-191
Dpsy_50876787_31-154
Bper_33592487_33-190
Bbro_33601549_33-190
Bper_33596147_50-207
Bper_33593267_32-160
Bbro_33600441_30-190
Gaul_39996135_20-167
Cvio_34101567_28-185
Bmal_53716046_23-159
Asp_56476344_659-780
Cvio_34105550_30-195
Hhal_18413619_35-164
Daro_53729407_19-172
Bpse_53717964_35-187
Sthe_51891769_26-180
Rpal_39936660_64-229
Peyr_46189181_155-289
Peyr_28870835_188-322
Peyr_28870271_154-292
Peyr_46187726_154-292
RspH_46192475_33-190
Rpal_39936811_33-193
Gaul_39997035_36-175
Xaxo_21108108_1-158
Xory_58582463_52-212
Pflu_48733017_24-183
Pasyr_23468688_24-184
Paer_15597769_24-180
Paer_46163700_24-180
Ccre_16124321_33-181
Atum_17937038_25-185
Bbac_42521886_28-185
Naro_48849248_17-176
Ctep_21674222_36-170
Ecar_50120271_32-193
Sent_62181664_35-185
Sent_56415093_35-185
Ecar_50121636_37-190
Ecar_50120444_30-192
Aper_14601157_25-163
Sone_24374574_120-230
Bpse_53722306_111-266
Blic_52005766_30-179
Bsub_16078921_1-152
Peyr_23470323_38-195
Peyr_28868542_37-194
Eamy_11127710_21-174
Esp_31580587_32-185
Peyr_26245281_21-174
Rsol_17430436_33-182
Agam_55246494_32-195
Vcho_9657614_34-195
Vcho_2190531_32-195
Ppro_54309471_32-195
Cvio_34103161_32-190
Msp_48833756_36-191
Lmon_16803266_528-691
Lmon_46907445_527-692
Lmon_47092025_527-692
Linn_16800258_527-692
Spla_2575805_29-182
Ppro_54302825_26-176
Vfis_59713239_30-184
Chut_48856658_27-180
Npun_23128582_207-361
Npun_23126935_207-360
Avar_53763807_207-358
Nsp_25530112_207-360
Nsp_17134035_32-183
Npun_23129859_32-181
Npun_23125110_31-184
Nsp_25532235_30-185
Avar_46134480_30-185
Psp_54031603_34-188
Mmag_46203087_30-181
Bpse_53722385_53-201
Bmal_53717169_36-187
Npun_23128359_30-183
Chut_48855250_26-181
Chut_48854073_18-168
Xaxo_21107507_34-181

DSLKEVENAEFAELVQGLHPLAAALKGD
DSLKEVENAEFAELVQGLHPLAAALKGD
SLYDEVLAGYDALAVGLALAPLHAALKWN
SLYDEVLAGYDALAVGLALAPLHAALKWN
SMETGLTSSFGSFAASLDEMAALERND
ALAEVLKTYQAYGLGTMPLNALGLQY
KMAAEVKSNEYQYMLGVQPMVDALRGD
GLSMNLVRRYRYSIDLBVDTMVEALREED
EKFANMKHKDKSLITVQLQFNDFDKFANMSK
KRIAFAEAVYKWTDLNTQVTELFDKGGL
KRIAFAEAVYKWTDLNTQVTELFDKGGL
SSLEKFLTTYNRWSDNDEIQKIMADAGS
FLVADVDRAGRLLETLSVHLRLVLSAP
QVYQETDRGRDLLALSAESMETVRRVS
GELNELKESTHQLYGLAEIAGLAK
TAFVDLKKGYNLQYGLVLELAEGLRT
ELINQVQLRLSLMAQLLPLALAK
SALEKDSROYNLITELFNQHSYQLQSVTA
PFAEEKPEWRDFKTN
PFIIEKPEWRDFNTN
PFIIEKPEWRDFNTN
DRVGAVETAFAHAYLEVAAKQMEAMGNQR
DRVGAVETAFAHAYLEVAAKQMEAMGNQR
DRVGAVETAFAHAYLEVAAKQMEAMGNQR
ELAATLQTRWRAYQSVLDELAAAVDAGQK
ELAATLQTRWRAYQSVLDELAAAVDAGQK
AAVKGAKELGKYRGKIDELVAARGSGF
QLLQGFVAAHQKMGDDYRKGLDAYKGA
GLIEQFASAAHAEKMG
RLSEGISGFRSQARILAEQAASARKG
GEFDMGLEGIAGYGLWGLYKOKIASGE
ALNRQHPRELLSRWTHDILPAINRAR
DRLHDVQSWRRIRLIVESLAGD
EQAERLAADWATLDRKGRIPAVDLNNGN
QHLEDAEALAGYQVLLQGMALRDBGN
ERLRSIGALTQYDGAITEVYGLKLAKAQ
AAVDAALVALKQYKVLVTSISDKMLK
IANDTALNALKQYKMLVSSISEMLKQ
PANDTALNALKQYKMLVTSISDKMLKQ
PANDTALNALKQYKMLVTSISDKMLKQ
KLIAAYADOKAAALNRRVFAALDAGD
EPFQKLKAAVAGYQKAKAVEMADGD
PFRDLVAGYERYQRATEIFAIGNGD
GELDAVTEALQAYIASADHVNAAREHA
KOLDAVTEALQAYIASADHVNAARETH
SALQKIPKSLDTYIASSEQIVGLALEN
AELSRIPKSLDTYIASSEQIVGLALED
QALVELRPDLLEYIAGAESIVGKALLD
QALVELRPDLLEYIAGAESIVGKALLD
AALLAVETPLNAYIAAEDIVGTADVD
EVIEGVEKPLLSYIDIAKKVVSGLANTD
RAIAETHKPEMERYIGQTRKIVGLASTQG
AARAAADPAFRAYVAIGREVSAAARE
ELVAREKGSNAEVQGL
ALVQAVSRDWSNLVQGVPEPLQKAAANT
DISREVLNNQALLEGVVPOMGLAQCGS
DISRAVLNNQALLEGVVPOMGLAQCGS
EADNMIASWTQLLDNGLTMMNLARDNRR
ATVDVAKNTWNTLSSIEPMNSALRND
SDVAVARELALITGLLEDLQARIGAL
ELLSHDBYHREK
OLINEVVTLEOYNQVNTQLQASQAYK
ANVEEIKENKAYNDLNOQVMQYASSL
SNIEDIKTSFTQYRALNRQQVITAYSS
MLTQQLQRYVKLRDWRHQSVICGD
MLTQQLQRYVKLRDWRHQSVICGD
SSFNGLISSYELMRKGMHQVQAVESGD
SSFNGLISSYELMRKGMHQVQAVESGD
SSFNGLISSYELMRKGMHQVQAVESGD
ASIKTLGEAYGRMKRGMHQVQAVESGD
ANAQQLLSQFEGMASNELAPMLQALBQ
EQARKLQARFEQVMVRELEPMLQAFANN
EQARKLQARFEQVMVRELEPMLQAFANN
KEIELEQIFEFKIVENELNPMLTAFENN
GSLQRLAEIRAYAKAFSDTVDLVEANN
LVSNQVDVETQHYITVLGVLTSGDGE
TSKLEEQFKKIGASQFAFAAHPEVRNDR
TSKLEEQFKKIGASQFAFAAHPEVRNDR
TSKLEEQFKKIGASQFAFAAHPEVRNDR
TSKLEEQFKKIGASQFAFAAHPEVRNDR
KHLDDLSTLSQKLEELQMTIELAKGN
TLKALEIDIDKCAELKQSIDLEFDDQ
QALKHIEKSMNLKFEELAATIDFVQDQ
KRIAMIQGLVQHMHTLDSCIKLREHDY
KQIATLEFLIAKLAELNOTIDLRQKSGS
NQFAALESIAAKFALLKQIYILYRQCG
QRLVILKPLITAKITELKQIKILRNQ
QRLVILKPLITAKITELKQIKILRNQ
ERITITKPLISRKLTLEQTIKRNDR
RRLDRLEPLITKLVILEQAINARSKD
NRLDLOPLITERMAMVMDVIELQEG
GYISRELEPLINERRAVMTVLDLQACG
GYISRELEPLINERRAVMTVLDLQACG
RRLDALPLVKRDLARETIDARIKG
RRLDALPLVKRDLARETIDARIKG
ORLATLRLMLEKKQHELAATIELRKTEG
ORLATLRLMLEKKQHELAATIELRKTEG
QWISIDLEPKITSRLNLOKEIYLRYNG
YIDTLQKTQVNRFFLNGLNLNTNANA
NNLDSVLKIVRYNITFENSINNDN
QHLAQORIVETRLRQGHVLDIYNSG

MAGYQSHAQNVLDALBGRFSGAIAAFDFWRASELLDAHEVAETRYR
MAGYQSHAQNVLDALBGRFSGAIAAFDFWRASELLDAHEVAETRYR
GLEANRLAERLPOAAERYIEAVDAPOGYAREGGRAVADAARVLER
GLEANRLAERLPOAAERYIEAVDAPOGYAREGGRAVADAARVLER
DAAYVLQKLVKAGQASAAFAARLGEFTNLDKLSSETLAHETRE
TDEYVEVLKOLGTLSEAFDLISNQFRQYAOQLSQBGLAQSERNE
VAAYYVLQDVLPLPSVKMDKVANEFRDFGLVEGNNMLTEASFAEDR
TDFYFMVNNYEGPRSAAFIAAGDEFSKYINEQOQMAETAAINTFDR
LFDVKNVNLTKVKYQENFMARESINQIEBVMDDYKNSIMAGSNANVR
DKEALNITLVTKGRDVRVGTETVDTVMVSNNNRMOAQCKSAEASRS
DKEALNITLVTKGRDVRVGTETVDTVMVSNNNRMOAQCKSAEASFARS
HQKALGLISSVGRDTRLEGDKALDEILTHQAOLECKKKSADALSST
EALERSIKERSRGEAGFAEAAQAFHKSSELDAKAEVROAIRTAQ
SPAALLEVKERFETAEKHFUDVDRALERKKVYRQGEVH
NNIDTFDVPDIOGQSDFTKRYHLYQESKSGSTAMDEQLSSLSAR
NQDIFFAVFPVQAYQSDFTKRYHLYQSDTDLAQKQHQQQLSSLDQAR
HLDQDPAFAFELEFPYQALSDPLPPLPSAALRHAYRQFIREKQ
ALDEAFNIDDAATMLDMDANNVEYSKASNTSESSLSITLMDY
KKKGWTSNQAVELSQQVLVRADIVNMLRK
KKKGWTSNQAVELSQQVLVRADIVNMLRK
KKKSAVNNQAVELSQQVLVRADIVNMLRK
LGEFVOLNPSAQRLNTAFDTAASAYLDRIPTDLDALVDDARASHLRAN
LGEFVOLNPSAQRLNTAFDTAASAYLDRIPTDLDALVDDARASHLRAN
LGEFVOLNPSAQRLNTAFDTAASAYLDRIPTDLDALVDDARASHLRAN
AEPALAAHMRQAQAEHAFQRMDEAFLARVQAHSDSEVRSGAEDTHVVAR
AEPALAAHMRQAQAEHAFQRMDEAFLARVQAHSDSEVRSGAEDTHVVAR
DVAALDRNLARGIDRPLEAAAVKEMEKAAQSGFGDSRAALLAASRRK
FKVGDQVAGMDPFPTELLTQASSELLAKRANANAAASADS
YRKGEFAFRAGFPSAGDQAVAGVDRAPVLEL
ARIHEMLKDTMEQAAVQASIDAEALQKQS
TAAANIGDAGIAPFPKPLYAARGALDKLSKLADDSAQOAEPTLETSMTART
AADDRGDPDPAPAFALVGDIDALVTLLEHNSSEKIK
ARSEVILDRFDLEPSPVMGINELVAMERSYAYDT
LSEAQVYVITQLDPVQVRKYGAAQLRQNLQEAQOAYDRARGVAV
ISQASAVLKNVDIPVTEQIVLDSQGVYLRQLVLAQAEATRAQOASAK
SGAASDQAKAKRDLVQANSVATQALGLMGEAVERAKAAKATKTQAADELFSAN
DTEQIRNDLQOQSVAATAARDDLAALQVITSAKKEQNT
QADQVRGNLQOQSVAATAARDDLAALQVITSAKKEQNT
DNQIDRTLRQOQLDLKSDAGLMACQVVSANKEKDSAVTS
DNQIDRTLRQOQLDLKSDAGLMACQVVSANKEKDSAVTS
VAGASRLLAGESRAGMKLANLETFRTLYKQDMTEAABEDRTTASI
AGSALLFIKSAKSSFTDIDSLVSDITLADNDSRORITARTGALSRG
PLAKKSDSVAARIASISALRRAILQRLAABDRREATV
DSAAAYAFNDKFGQDQTRMKAISERILALNBSASRQAEQDSHR
ADSTAAAYQFNDITFGQDQTRMKAISERILALNBSASRQAEQDSHR
PESAOQHLGTFSTAFSQLEQMAALSELIBTNTRQTSQETAAIHNA
PDRAQOELGTFSFAFTLEGQMSITLSDLEBNSKASGERTQRTTSAN
PVAARAEPLQFVQAFKLEGRNLSLLEKHKVQOTNRAREDSM
PVAARAEPLQFVQAFKLEGRNLSLLEKHKVQOTNRAREDSM
PAAAHAKLTDGQOQSTLEDKMAEASTIEAASHDNTAAKSLG
PDAAVKLVDPDFVROPSALETTMEKASQIDALSSATLAESEKASSTIE
YRAALSELEDFNKSFVKLEEMALGLGLEKDAEAAHGSAGGLRT
GAVPQDGAELQRFQHLFTQLEALMSKISDAVEAHSSETVAAESSAAQAR
NAGLESAPQHNDKALSLNSGADLMHIDNDNTLTA
LDDYQNAQKDVPPSLSRQFASLGFASNASAEKFDAGAVFEQITLVGK
LTAWSEHASTVTPALSRAFASAEERFHEAGAMLDNTRVIMVGD
LTAWSEHASTVTPALSRAFASAEERFHEAGAMLDNTRVIMVGD
CEEFQRLFRKAYPPLSAVFGSSMDQYAAIISSTSMKRETVLVI
PEAFQRIERSVYPPVSLTFGGDIKRYSDGITSLSIPVNEHNDKRN
SGAGAEGLRDALQORLARSLEABERDEVYAS
AKIMOPIDQEFALDKRTDAQMDNRLDSLHMAINSQ
MLBRADLAKNAMKLEGLDETQYSGYDSNAEKFKQEELNSKRYNDE
KKAESTHFGERTLRKVDLPSVNLQVLESLEQVEDLKAETRANGTMS
KKAESTHFGERTLRKVGAPVAPVNLNRQVDELDKDEIQNGKMSQ
TAAATNINRTVEKPPAEQVKTTLQTLREQCKQRAGEERFDAQESSATQAN
TAAATNINRTVEKPPAEQVKTTLQTLREQCKQRAGEERFDAQESSATQAN
IKKASEINRNEVKAYADKTFGALSTLKKHQDAVENKVVQAKRKDFRA
IKKASEINRNEVKAYADKTFGALSTLKKHQDAVENKVVQAKRKDFRA
IKKASEINRNEVKAYADKTFGALSTLKKHQDAVENKVVQAKRKDFRA
IKKASEINRNEVKAYADKTFGALSTLKKHQDAVENKVVQAKRKDFRA
MEARAGTYRDRTYTYGTMKRDNALLDGLITQQAQRQLHQEQSYASGR
MTTAQNIYRDKYAPTYGEMRKQANQILDTLLQQAQRQNHASVESFEAGR
MTTAQNIYRDKYAPTYGEMRKQANQILDTLLQQAQRQNHASVESFEAGR
IVTAKQIYHERYTKSYGEMRKRAQSIMNSMLEQAREQNEVSQKSYSSG
PEALAKOYGGQTRPALQALLAATAEASAAQOQHMMAEMSGSLVIMSATR
ERALLLESBTGPALAEIHAIFLDITTEENHMQGFREALITQNR
NDRFQKLAADIQGLETAGNELKTSAGQQAALQAKMKELNGCIQQLATAMNQLNEQSKQIS
NDRFQKLAADIQGLETAGNELKTSAGQQAALQAKMKELNGCIQQLATAMNQLNEQSKQIS
NDRFQKLAADIQGLETAGNELKTSAGQQAALQAKMKELNGCIQQLATAMNQLNEQSKQIS
NDRFQKLAADIQGLETAGNELKTSAGQQAALQAKMKELNGCIQQLATAMNQLNEQSKQIS
OTATELVKTHGKNLMNEIRQLDTMKDEERLEFIRSADATELIRNA
TAALQVLSDVGRQYMDNIRWYLLSFSSEENRLLEKRGDYREVR
LPKALSIVKEDKGRAYMDLREDLNFTNIEILLLEORKGDFKENRSQ
DKAQSVASENGKRMNDKIRDNIELAKDEEFRLLEIRKDKSDNDARD
EALLOVLTNKSQNLMDDIRKAIDEIENEERAGORQOLSQTAQITWRKTT
FEAALQVITQNHGKIMDDIRKVIHEIENEERAGORQOLSQTAQITWRKTT
OTALQVITQNHGKIMDDIRKVIHEIENEERAGORQOLSQTAQITWRKTT
FOALQVITQNHGKIMDDIRKVIHEIENEERAGORQOLSQTAQITWRKTT
EKEAALKIVOTEEQOLNCKIRVIAEENENLLQORSIQTEAYRT
LESALQIVTDRGQIMEDIRMLTAMNEENELLKORSEANASAHQ
EASQKALIDQGLMDQIQKQIMKTEENELLKORSEANASAHQ
FEAAQKSISSDRGIMAQIRISISQMAEELKLLORTAMANAQANNTL
FEAAQKSISSDRGIMAQIRISISQMAEELKLLORTAMANAQANNTL
EAGVEIKITGKRTLTSEIRKVDMEEREEDKLLKORAEANQANDAR
LETALRIVTIDGENDMREANVLAAMLADEQRLDLARLQANDAR
DAAQAVISDAGITFMDRARMVNMEMATVQOOLRLTRDRARATRO
DAAQAVISDAGITFMDRARMVNMEMATVQOOLRLTRDRARATRO
EAVRQKILSDKDCSGKEITQOILHDSLEENLLQOQMSQSSQK
QAEVPLQKGAESTHEIYNLVKKRNNRGLLQORNEKEADKPTP
NAKRFNSKILGAVMRTIRSQNOMILENIYLYVEREYENESFLTP
LEPRAADROAFRFTSAIPRELAMVQREQLLQARQSSQSAH

Figure B.3 continued

Xory 58581502 34-181
Xcam 21203751 34-180
Xaxo 211077664 31-181
Xcam 21230901 31-183
Rpal 39935361 37-185
Rbal 32436139 63-213
Npun 53687340 14-163
Lpne 52840485 46-200
Lpne 54293235 44-201
Psp 54028600 33-179
Psp 54030014 27-182
Npun 23129372 36-186
Mdug 48864300 37-191
Bsu1 23348515 37-185
Rrut 48763304 34-183
Atom 15889792 34-185
Mlot 13474773 16-17
Rsol 1084043 66-207
Rsol 17427298 25-180
Rnet 53761798 27-182
Rmet 48770512 28-183
Rgel 47574299 34-184
Psp 54032248 52-203
Psp 54032135 25-175
Bbac 45224788 28-184
Rsph 46192383 29-182
Nro 48850773 37-186
Drad 15897571 51-199
Rpal 39935500 31-186
Bjap 27379400 31-186
Bjap 27379387 1-124
Pflu 48731493 44-184
Psp 46188881 30-168
Psp 28869899 43-182
Pput 24985348 43-179
Bfun 48785827 44-182
Mfla 45520823 45-187
Ppro 54308182 32-133
Ppro 54302047 35-187
Xory 58582467 55-214
Xory 21108104 4-133
Pspyr 23468728 25-194
Bcep 46321500 33-168
Rrub 48766644 32-188
Mmag 46204782 17-166
Dpsy 50875181 214-367
Dpsy 50875181 53-203
Psp 54308182 35-175
Lint 24214048 27-178
Lint 45658825 27-178
Bjap 27380803 69-216
Dhaf 23116533 73-221
Xaxo 21106794 50-198
Xory 58583559 50-198
Xcam 21232865 34-184
Pspyr 23468998 40-186
Pspyr 28869332 13-157
Xaxo 21107031 39-186
Xcam 21230299 39-186
Mfla 46210474 48-205
Vvul 27359643 30-187
Vvul 3779272 30-187
Nfar 54022117 37-190
Soc 6562854 30-186
Mavi 41409388 37-192
Tfus 48853038 48-200
Rxy1 45548044 36-189
Gvio 37520854 36-203
Ppro 54302045 49-202
Ccre 16217375 29-179
Sthe 51892221 47-198
Sthe 51892220 31-185
Sthe 51892086 31-179
Bjap 27377458 27-183
Bjap 27377456 27-182
Rpal 39933216 26-159
Rrub 48764325 9-127
Rsp 36958591 29-192
Bjap 27377617 93-243
Bjap 27377618 16-163
Bhal 10173491 26-176
Rrut 48763402 42-132
Sthe 51891861 43-195
Vpar 28900441 195-324
Psp 54303525 181-310
Msp 48832744 193-326
Bjap 27375494 35-174
Mmag 23014424 13-172
Mmag 46201214 26-180
Mmag 23014424 180-325
Bjap 27375494 184-330
Daro 41273787 35-182
Tden 52007874 16-169
Wsuc 34557595 31-176
Msp 48833622 31-186
Mmag 23016728 31-186
Asp 56475846 38-186
Gmet 48846495 29-184
Mmag 23014950 31-187
Cthe 48860417 33-191
Wsuc 34557782 18-193

Figure B.3 continued

Jsp_28201217_24-165
Gsul_39996242_32-184
Msp_48831171_31-189
Cvio_34101840_32-198
Bfun_48788236_15-180
Bfun_48783867_31-180
Gsul_39998082_27-169
Vvul_37679586_11-184
Vvul_27362348_9-182
Gmet_48844217_30-202
Gsul_39997468_30-202
Gsul_39995872_33-218
Cups_57505871_291-429
Cjej_57238206_291-435
Ccol_57168164_291-432
Dvul_46579163_32-189
Ddes_53691502_33-184
Wsuc_34558440_32-190
Ssp_52012288_26-183
Atum_17936772_24-180
Selo_56750264_62-198
Selo_46129799_62-198
Bcer_30019279_33-192
Bant_21399048_6-166
Mmag_23016539_32-190
Save_29609038_91-240
Scoe_6855387_34-184
Krad_53768261_11-176
Mmag_23010448_105-259
Bfun_48780983_34-191
Bjap_27377964_44-217
Lint_24216205_37-185
Wsuc_34557237_30-187
Wsuc_34557419_38-224

CHISQLDTEALSYPEKVKQITLADVEG-----AERTISTNIEGLVQSAWAEKMDLVALQFSSASDILAASVVDOR--
KELDDVEAAFTAFHLSGVKMAKVMYMG-----TGAGNPLMKEFFDNAHEVLIEKVEKLQKRSQVDEALGNSRDNVAAGVH--
AQIKKLSSALTAWYAVGKKMAQGYIEGG-----FVLGNQLMGGFDEQAEILQGGQLKPFLLDGGQQQIYTLTMMVDEVNYPE--
SOLDSVQR F AFMERDRDIAMKRSGD-----AAQRDKALVAGIEQIFTHISQNTNALVSIITLDMKKVQAQSNISYGVIVAR--
ERVVQVRQAVQAYRCWSRVDMLASGRS-----KQVLARFRDLAPTDVLDLRFQQLAAFGQAEERRLDRLSNAAQAAERQQ--
ERVHEMORAVEAYRAQTVAISQALHAG S-----ENSFAAQESGALPROVALFRDELAAFGDEASRLDTERSAALARRARQQ--
A GVAGED LDELVQEVQTRYASVVR-----TAWEFFELQDAGRGDEAMALFRGR EYFDRBEVL VDK LDD-----
RLEKIESSMKQFTDAFEAQILAKNGGD-----IDESLVALRTLYQTHLPIENMLDEASEEELAGADAAMGALDGLSLNIE--
RLEKIESSMKQFTDAFEAQILAKNGD-----IDESLVALRTLYQTHLPIENMLDEASEEELAGADAAMGALDGLSLNIE--
SHAETLTETWGEYKVAEKIAYKTAVIAGQOT F-----STLSESRLLSELSGASEFVARDIDDLIETVKGMMKDAGKETSRVKTSS--
IRPDSVDTVMEEHAKALLDTNAEYKVDRLIAYKSGVLGSGSVSESVIVETRLISELSGASEFVARDIDDLIETVKGMMQVQGTQRIAS--
TYAAAVLASNGEPEKAADQLIAHKQRLKGLRSGVVDQAAKDALADETLNRLALETITRDTSENAKLDIDDLADLYLESOTYAGLKTETTRQSTR--
AMLESQEKVFNGFKTSLFEGMKKFKEVY-----EEEKNDDEKIELMDEKFKELNALDNETKEVIAKI--
AMLESQEKVFEGFKASLFEQMKKFKEVY-----EEEKSIDAKIKMMDELKEMKALDEETSQMMSKLGDENQ--
AILESQEKVFEGFKTSLFEGMKKFKEVY-----EEEKSIDAKIKMMDKLQSKIKNLNEETNEVMAKIKI---
ALIAEASRSTLAFASGLAFVERLAAVP-----TERHTHYFAIRRT LAEASRRTFNRIVESKDAHETANTAFGDELDSAR--
OLVAKAKRAALAFVQDGLRFATMQSIF-----EGMRHLQIAYEGSATPLAQSKRKYPGELTITIKKAYDNEVYRFKDTI--
QALEHSRSQVLAFIDHGKQVQS KEDPS-----DEHRMALYAEYRTA PLATASRESFQKLVLDKNNABEASQKELDSTLTLLD--
AFINRIADIDALQEMERAAIKAYQAGN-----IDRARGLLFGGAHYQAHKLIDVVFQFRMVQVDRKALDELTDAHQRSQWFE--
TSFEETIKKSADVLDRMEREAI GFYRSRG-----SEAQQVLDGEYYRLHRLGLDVTETLRATLATRTATVQA SRSSD--
KEIAALSAIQSKLEEQONLMNAKQEAT-----PSLDASYYNVHSQINQIVQNERILLDRVINVLVYR--
ELFQMKKSLPSYKDTYKKLFKAKATN-----PSLDASYYNVHSQINQIVQNERILLDRVINVLVYR--
ELFQEMKSLPIYMGTYKKLFKAKVKN-----ESQMVNEFQKELKPRGIELAGYVLDLESYVSNLAQOVFKDSQKMDRS--
LLVESYKPALRFYDQVFGTYFPALG GD-----ESQMVNEFQKELKPRGVELAGYVLDLESYVSNLAQOVFKDSQKMDRS--
DTIAKLNKLLPEYKGLIERARA NRQGF-----LGAAYLRYANDTMOQKMLPAADQLYTKENQRLSRSDYD ATP-----
DTIAKLNKLLPEYKGLIERART NRQGF-----VGGAYLRYANEMKQKMLPAADLYTKENKRLGADYDGAQ-----
AVVAGVEEDVHAWFEALRAEYALATD-----DAVAVSFGANRDKARAYEGLAEETERAGTALAAGQDFAATVRRGQ--
AKMATLQMAAGITHELNQPLAALRGRL-----ENAGAFIRQGREEEAANLTRISVLDRLGKLTGQLRSFSRRSGTE--
DAFRNVLMKMQADAWPAEANRLVRIG-----HRDEAFQMLTAQITPTFARIQKWLGEERARQGEOLGGQVSEIERALQWTR--
VDLARMAGVLKYMGRSLEDPLAVRDATLAALISG-----RRDEARTISRGFAKALAGPLDLSQIRLEADITDRS RKL LARQLNIT--
ENLKLEDDGINKFAQVATLTHLGLKDNR-----ROEAOILYVRGVNPLSASLEQTVKTLTFEASMSQKNEEDAAESQ--
DLLLRA RQIGDSLQRIIDIAEVLASKE-----LARI TLSYERILETIEKTRLSLLGALIERENKNAYQKTESVNRNDEETI--
PLWPEPSLSERRAMIS TEEKILLGROISEEVFALLALH X-----ESEATQKIR ELVPSINIVNTYLSQILRHKNLSAAIKRSTINLYQTTU--

Agam, *Anopheles gambiae*; Aper, *Aeropyrum pernix*; Asp, *Azoarcus* sp; Atum, *Agrobacterium tumefaciens*; Avar, *Anabaena variabilis*; Avin, *Aerobacter vinelandii*; Bagr, *Brevibacillus agri*; Bant, *Bacillus anthracis*; Bbac, *Bdellovibrio bacteriovorus*; Bbro, *Bordetella bronchiseptica*; Bcep, *Burkholderia cepacia*; Bcer, *Bacillus cereus*; Bfun, *Burkholderia fungorum*; Bhal, *Bacillus halodurans*; Bjap, *Bradyrhizobium japonicum*; Blic, *Bacillus licheniformis*; Bmal, *Burkholderia mallei*; Bpar, *Bordetella parapertussis*; Bper, *Bordetella pertussis*; Bpse, *Burkholderia pseudomallei*; Bsub, *Bacillus subtilis*; Bsui, *Brucella suis*; Bthu, *Bacillus thuringiensis*; Cace, *Clostridium acetobutylicum*; Caur, *Chloroflexus aurantiacus*; Ccol, *Campylobacter coli*; Ccre, *Caulobacter crescentus*; Ccjp, *Corynebacterium diphtheriae*; Ceff, *Corynebacterium efficiens*; Cglu, *Corynebacterium glutamicum*; Chut, *Cytophaga hutchinsonii*; Cjej, *Campylobacter jejuni*; Ctep, *Chlorobium tepidum*; Ctet, *Clostridium tetani*; Cthe, *Clostridium thermocellum*; Cups, *Campylobacter upsaliensis*; Cvio, *Chromobacterium violaceum*; Daro, *Dechloromonas aromatica*; Ddes, *Desulfovibrio desulfuricans*; Ddig, *Desulfovibrio gigas*; Dhaf, *Desulfotobacterium hafniense*; Dpsy, *Desulfotalea psychrophila*; Drad, *Deinococcus radiodurans*; Dvul, *Desulfovibrio vulgaris*; Eaer, *Enterobacter aerogenes*; Eamy, *Erwinia amylovora*; Ecar, *Erwinia carotovora*; Echr, *Erwinia chrysanthemi*; Ecol, *Escherichia coli*; Epyr, *Erwinia pyrifoliae*; Esp, *Erwinia* sp; Gkav, *Geobacillus kaustophilus*; Gmet, *Geobacter metallireducens*; Gsul, *Geobacter sulfurreducens*; Gvio, *Gloeobacter violaceus*; Hhal, *Halomonas halodentrificans*; Hhep, *Helicobacter hepaticus*; Iloi, *Idiomarina loihiensis*; Jsp, *Janthinobacterium* sp; Krad, *Kineococcus radiotolerans*; Linn, *Listeria innocua*; Lint, *Leptospira interrogans*; Lmon, *Listeria monocytogenes*; Lpne, *Legionella pneumophila*; Mace, *Methanosarcina acetivorans*; Mavi, *Mycobacterium avium*; Mbar, *Methanosarcina barkeri*; Mbov, *Mycobacterium bovis*; Mdeg, *Microbulbifer degradans*; Mfla, *Methylobacillus flagellatus*; Mgal, *Mycoplasma gallisepticum*; Mlep, *Mycobacterium leprae*; Mlot, *Mesorhizobium loti*; Mmag, *Magnetospirillum magnetotacticum*; Mmaz, *Methanosarcina mazei*; Msp, *Magnetococcus* sp; Mtub, *Mycobacterium tuberculosis*; Naro, *Novosphingobium aromaticivorans*; Neut, *Nitrosomonas europaea*; Nfar, *Nocardia farcinica*; Npun, *Nostoc punctiforme*; Nro, *Nostoc* sp; Paer, *Pseudomonas aeruginosa*; Pflu, *Pseudomonas fluorescens*; Plum, *Photobacterium luminescens*; Pmul, *Pasteurella multocida*; Ppro, *Photobacterium profundum*; Pput, *Pseudomonas putida*; Pres, *Pseudomonas resinovorans*; Psp, *Polaromonas* sp; Psyr, *Pseudomonas syringae*; Raqu, *Rahnella aquatilis*; Rbal, *Rhodopirellula baltica*; Rcen, *Rhodospirillum centenum*; Retl, *Rhizobium etli*; Reut, *Ralstonia eutropha*; Rgel, *Rubrivivax gelatinosus*; Rleg, *Rhizobium leguminosarum*; Rmet, *Ralstonia metallidurans*; Rpal, *Rhodospseudomonas palustris*; Rrub, *Rhodospirillum rubrum*; Rsol, *Ralstonia solanacearum*; Rsp, *Rhizobium* sp; Rsph, *Rhodobacter sphaeroides*; Rxyl, *Rubrobacter xylanophilus*; Save, *Streptomyces avermitilis*; Scoe, *Streptomyces coelicolor*; Selo, *Synechococcus elongatus*; Sent, *Salmonella enterica*; Sfle, *Shigella flexneri*; Smel, *Sinorhizobium meliloti*; Sone, *Shewanella oneidensis*; Spia, *Spirulina platensis*; Ssp, *Silicibacter* sp; Sthe, *Symbiobacterium thermophilum*; Styp, *Salmonella typhimurium*; Tden, *Thiobacillus denitrificans*; Tfus, *Thermobifida fusca*; Vcho, *Vibrio cholerae*; Vfis, *Vibrio fischeri*; Vpar, *Vibrio parahaemolyticus*; Vvul, *Vibrio vulnificus*; Wsuc, *Wolinella succinogenes*; Xaxo, *Xanthomonas axonopodis*; Xcam, *Xanthomonas campestris*; Xory, *Xanthomonas oryzae*; Ypes, *Yersinia pestis*; Ypse, *Yersinia pseudotuberculosis*; Zmob, *Zymomonas mobilis*

Figure B.3 continued

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 6

PAS_Aer

2	Daro_8-120	-----TDVETRLPEGOFTL	YSRT	DLKGV	ITEANEAFACISAYR	REEL	LG	-----NNNM	-----	VRHPDMP	AAA	FA	-----	DMWNL	
2	Daro_4-120	-----NLPVTN	YETHLPEGEFTL	YSST	DLQGNLVEAN	AEAKISNFS	REEL	IGQ	-----PHNM	-----	VRHPDMP	AAA	FA	-----	DMWNL
3	Reut_6-119	-----TD	EYRLPSDEVTL	ITRT	DAQGNLEYAN	AEFRSSGYDRAS	ELIGQ	-----PONI	-----	VRHPDMP	AAA	FA	-----	DMWNL	
4	Rgel_11-124	-----	AGLVSAYVQAQPLIL	TL	T	DLQGAISFANKAEL	LOTGYMAQVLGA	-----PHST	-----	VRHPDMP	PKV	FA	-----	DMWNL	
5	Bcep_8-121	-----	TQREFEPDDATL	MSST	DANSYITQYAN	AETQVSGFS	PEEL	EGQ	-----PHNV	-----	VRHPDMP	KEA	FA	-----	DMWNL
6	Bcep_8-121	-----	TQREFEPDDATL	MSST	DANSYITQYAN	AETQVSGFS	PEEL	EGQ	-----PHNV	-----	VRHPDMP	KEA	FA	-----	DMWNL
7	Bfun_8-121	-----	TQREFEPDDATL	MSST	DTESYITQYAN	AETQVSGFS	PEEL	EGQ	-----PHNV	-----	VRHPDMP	KEA	FA	-----	DMWNL
8	Rsol_8-121	-----	TQREFEPDDATL	MSST	DTQSYITAYAN	AETQVSGFS	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
9	Bcep_8-121	-----	TQREFEPDDATL	MSST	DADSITQYANTT	FAVSGFT	PEEL	EGQ	-----PHNA	-----	VRHPDMP	KEA	FA	-----	DMWNL
10	Bcep_6-121	-----	PVTQREFEPDDATL	MSST	DADSITQYANTT	FAVSGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
11	Bmal_1-101	-----	TQREFEPDDATL	MSST	DPHGRITQYAN	TFVHVSGFS	PEEL	EGQ	-----PHNV	-----	VRHPDMP	KEA	FA	-----	DMWNL
12	Bpse_8-121	-----	TQREFEPDDATL	MSST	DPHGRITQYAN	TFVHVSGFS	PEEL	EGQ	-----PHNV	-----	VRHPDMP	KEA	FA	-----	DMWNL
13	Bmal_8-121	-----	TQREFEPDDATL	MSST	DPQSVITQYAN	AETQVSGFS	PEEL	EGQ	-----PHNV	-----	VRHPDMP	KEA	FA	-----	DMWNL
14	Ecol_8-121	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNM	-----	VRHPDMP	KEA	FA	-----	DMWNL
15	Ecol_8-121	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNM	-----	VRHPDMP	KEA	FA	-----	DMWNL
16	Styp_8-121	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL
17	Styp_8-121	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL
18	Ecar_8-120	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL
19	Ypes_8-120	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
20	Ecar_8-121	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNM	-----	VRHPDMP	KEA	FA	-----	DMWNL
21	Ecar_8-118	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
22	Reut_8-121	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
23	Rmet_8-121	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
24	Avin_10-121	-----	TQREFEPDDATL	MSST	DLQSYITHAN	TFVQVSGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
25	Rsol_8-121	-----	TQREFEPDDATL	MSST	DPAGNITVFNAPFV	IRISGFT	PEEL	EGQ	-----PHNV	-----	VRHPDMP	KEA	FA	-----	DMWNL
26	Bbro_8-121	-----	TQREFEPDDATL	MSST	DTKGRITVFN	AFVSGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
27	Tden_8-121	-----	TQREFEPDDATL	MSST	DREGITVFN	AFVSGFT	PEEL	EGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL
28	Bjap_8-121	-----	TQREFEPDDATL	MSST	DLKGRITVFN	DFLAAAGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
29	Rpal_8-121	-----	TQREFEPDDATL	MSST	DVKGRITVFN	QFVKASGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
30	Daro_8-121	-----	TQREFEPDDATL	MSST	DLKGRITVFN	DFLDSISGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
31	Neur_4-121	-----NM	TQREFEPDDATL	MSST	TAKGVITYINEP	FIRMSGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
32	Mfla_8-121	-----	TQREFEPDDATL	MSST	DLQGRITYINODE	FIEVSGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
33	Tden_8-121	-----	TQREFEPDDATL	MSST	DLRGNITVYNODE	FVDSGFT	PEEL	EGQ	-----PONI	-----	VRHPDMP	KEA	FA	-----	DMWNL
34	Rgel_8-120	-----	TQREFEPDDATL	MSST	DLKGRITYCN	AFITVSGYARE	ELIGQ	-----PHNM	-----	VRHPDMP	KEA	FA	-----	DMWNL	
35	Bfun_4-120	-----NQPV	TQREFEPDDATL	MSST	DLTGRITQYCN	PAFIAVSGFT	PEEL	EGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL
36	Bbro_8-121	-----	TQREFEPDDATL	MSST	DLKGRITYCN	PAFIAISGFT	PEEL	EGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL
37	Lint_8-121	-----	TQREFEPDDATL	MSST	DMKGRISV	QDFANISGFT	PEEL	EGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL
38	Psyr_8-121	-----	TQREFEPDDATL	MSST	DTGRITVYCNDAF	VDSIGYSAE	ELIGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL	
39	Psyr_8-121	-----	TQREFEPDDATL	MSST	DTGRITVYCNDAF	VDSIGYSAE	ELIGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL	
40	Pput_8-121	-----	TQREFEPDDATL	MSST	DTGRITVYCNDAF	VDSIGYSAE	ELIGQ	-----PHNI	-----	VRHPDMP	KEA	FA	-----	DMWNL	
41	Pflu_8-121	-----	TQREFEPDDATL	MSST	DAKGVITYCNDAF	VEISGFT	PEEL	EGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL
42	Paer_8-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
43	Psyr_8-122	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
44	Psyr_8-125	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
45	Pput_8-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
46	Pput_8-120	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
47	Pput_8-120	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
48	Vvul_8-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
49	Vcho_8-120	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
50	Vpar_4-117	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
51	Ppro_1-115	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
52	Vvul_5-119	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
53	Sone_6-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
54	Sthe_8-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
55	Sone_14-127	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
56	Ecol_6-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
57	Vcho_11-124	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
58	Pflu_6-123	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
59	Vpar_8-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
60	Vvul_8-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
61	Vcho_24-136	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
62	Ppro_6-119	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
63	Ppro_8-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
64	Sone_6-119	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
65	Rpal_7-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
66	Rpal_6-120	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
67	Rpal_7-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
68	Mmag_1-109	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
69	Bjap_7-122	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
70	Rpal_1-124	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
71	Mmag_20-133	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
72	Mmag_8-122	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
73	Wsuc_9-123	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
74	Wsuc_7-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
75	Wsuc_7-121	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
76	Daro_1-109	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
77	Cjej_1-113	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	
78	Cvio_9-125	-----	TQREFEPDDATL	MSST	DLKGRITYCNDAF	VDSIGYSAE	ELIGQ	-----PHNL	-----	VRHPDMP	KEA	FA	-----	DMWNL	

PAS_Che

79	Ypes_15-131	-----PRAELTSIDNAVE	MIIF	KPDGT	VQVNNLFLAAMGYQKDE	VIGK	-----HHKI	-----	FCDPQYA	SDAYR	-----	RHWQL
80	Ypes_1-112	-----MTSI	NAVE	MIIF	KPDGT	VQVNNLFLAAMGYQKDE	VIGK	-----HHKI	-----	FCDPQYASDAYR	-----	RHWQL
81	Ypes_11-127	-----PRAELTSIDNAVE	MIIF	KPDGT	VQVNNLFLAAMGYQKDE	VIGK	-----HHKI	-----	FCDPQYA	SDAYR	-----	RHWQL
82	Paer_17-134	-----EDVAVR	LEAIGQNV	TIRF	TPDG	ILSANLFLAAGYSADE	VIGK	-----HHKI	-----	FCEEDYQASAAV	-----	RHWQL

Figure C.1 Unedited alignment of chemotaxis PAS domains constructed using PCMA and ClustalW, and visualized with VISSA.

83 Paer_17-134 -----EDLAWR--LDAIGQNV--TIRF--TPDGQILSANLFLAVVGYSADEIVGK-----H-RI-----FCEEDFQASAAVY-----RFWKE
84 Vpar_13-130 -----S-EYRRFIKGLSDSMA-MIEF-DTRGILLNANDLFLSCVGYTREATIVGK-----HHSI-----FCDRGVQVSPRYQ-----QFWDG
85 Vcho_53-169 -----S-HODNIVQSLSEHIA-MIEF-DTSGVILSANLFLNAGVYRLLEITGK-----H-RI-----FCSPAECQSOQYQ-----QFMTS
86 Vcho_23-140 -----N-DSDIVSALKRNIA-MIEF-DPQGIILSANLFLSAMGYTKEEVIGQ-----HHSI-----FDPETVNSLEYA-----QFWSK
87 Pput_23-138 -----A-CAKLAALGRSMA-MIEF-APDGTILSANLFLCQMGYSABEIVGK-----H-RI-----FCEPDYARSAEYQ-----QLWRE
88 Pflu_23-138 -----A-CAKLAALGRSMA-MIEF-TPEGIVLDANLFLCKMTGYSAEIVGK-----H-RI-----FCEAFYRSEYQ-----KLWRE
89 Agam_1-111 -----A-VEAVLEGIA-TIME-TPDGEILSANLFLSLMGYSABEIVGQ-----H-M-----FCEPELARSADYQ-----QFMAQ
90 Pput_144-265 -----R-EHESILKALMRSTA-MIEF-DLDGILLTANDNRLATGYRLLEIVGK-----HHRM-----FCEPQVNSSEYQ-----AFWER
91 Pflu_140-257 -----R-EHENLIGALVRSTA-MIEF-DLSCNVLSANDNRLQMGYSABEIVGK-----H-RQ-----FCEPEEYNSAEYQ-----NFWER
92 Pflu_145-261 -----R-ENSAFTQALLRSTA-MIEF-DLSCNVLTANDNRLQMGYSABEIVGK-----H-SM-----FCDPAETQSSEYQ-----EFWAN
93 Psyr_170-286 -----R-ENSAFTQALLRSTA-MIEF-DLSCNVLTANDNRLQMGYSABEIVGK-----H-L-----FCDPAETQSSEYQ-----EFWAN
94 Psyr_140-257 -----R-ENEALVNALQRSTA-MIEF-TLDGTVLTANDNRLAMGYELNLEIVGK-----HHKM-----FCVPPEESADAYT-----QFWER
95 Psyr_155-272 -----R-ENEALVNALQRSTA-MIEF-TLDGTVLTANDNRLAMGYELNLEIVGK-----HHKM-----FVPEESNADAYS-----HFWER
96 Vpar_141-258 -----R-EYEDMLNALISMA-MIEF-TLDGTVLTANDNRLSTMGYKHEIVGK-----H-RI-----FCLPEEANSSEYQ-----QFWKE
97 Psyr_125-242 -----Q-ESIELLCAINRSTA-MIEF-SLDGVLNANLFLRIETMGYSABEIVGK-----H-RI-----FCLDEDAASAEYQ-----KFWKE
98 Psyr_140-257 -----A-ENEALDIALERSTA-MIEF-NLDGTVLTANDNRLQMGYSABEIVGK-----H-SM-----FCHAGDABSPQYT-----AFWAK
99 Pput_144-260 -----H-EHESILKALSRMA-MIEF-TPQGVILKANLFLDTMGYRLLEIVGK-----H-GL-----F LAHERSQAQYR-----EFWKS
100 Pflu_143-260 -----K-EESMLAALGRSMA-MIEF-TPEGNVLTANDNRLKMTGYSADEIVGK-----HHSI-----FCHRVFASQAQYR-----AFWAK
101 Pflu_144-257 -----S-SKLAAVRAMA-MIEF-EPNGVILKANLFLNVMGYALABEIVGK-----HHS------FCEPTLVNSPEYT-----EFWRK
102 Psyr_143-257 -----S-SKLAAVRAMA-MIEF-EPNGVILKANLFLNVMGYALABEIVGK-----H-RS-----FCEPVLNSPEYQ-----EFWRK
103 Paer_140-257 -----H-EMQSKLDALSRMA-MIEF-DLDGVLNANDNRLATMGYRABEIVGK-----HNRQ-----FCEPQVNRGPQYQ-----DLWRE
104 Paer_117-234 -----H-EMQSKLDALSRMA-MIEF-DLDGVLNANDNRLATMGYRABEIVGK-----HNRQ-----FCEPQVNRGPQYQ-----DLWRE
105 Pflu_144-257 -----S-AKLQAIIDRAMA-MIEF-DLDGVLNANDNRLATMGYRABEIVGK-----H-RL-----FPAQVNSSEYQ-----DLWRE
106 Bjp_134-256 -----K-KVVRAN--TEASKVSAISRQA-MIEF-KLDGTVLTANDNRLQMGYSABEIVGK-----HHSI-----FVQASERDGGAYR-----EFWAA
107 Rpal_135-250 -----S-DAGKLAALGRAQA-MIEF-AMDGTVLTANDNRLAAMEYRLEIVGK-----H-M-----FVPEPSVRSSEYQ-----EFWAR
108 Bjp_263-378 -----S-LAGQIAALDKAQA-MIEF-NMDGTVLTANDNRLQMGYSABEIVGK-----HHSM-----FVPEPAERDGGAYR-----EFWAA
109 Rpal_256-371 -----S-MAGQIAALGRAQA-MIEF-NMDGTVLTANDNRLQMGYSABEIVGK-----HHSM-----FVPEPAERDGGAYR-----EFWAA
110 Rpal_255-372 -----A-FSGQIDALIRKSQA-MIEF-SIDGTVLTANDNRLQMGYSABEIVGK-----HHSM-----FIDPAERDGGAYR-----EFWAA
111 Rpal_133-249 -----S-DFAGQVAAA-RSQA-MIEF-NMDGTVLTANDNRLQMGYSABEIVGK-----HHSM-----FVPEPAERDGGAYR-----EFWAA
112 Bsp_90-205 -----S-LLGKVNALIRKSQA-MIEF-KLDGTVLTANDNRLQMGYSABEIVGK-----HHSM-----FVPEPAERDGGAYR-----QFWET
113 Rpal_11-128 -----D-LAAAMLAALIRKSQA-MIEF-DLDGVLNANDNRLQMGYSABEIVGK-----HHRM-----FVPEPSHDSSEYQ-----EFWTA
114 Bjp_19-134 -----S-ADAQLAALIRKSQA-MIEF-AMDGTVLTANDNRLQMGYSABEIVGK-----KHAM-----FVPADQDRSSEYQ-----AFWAK
115 Rpal_90-205 -----S-LOEKVAAIRKSQA-MIEF-AMDGTVLTANDNRLQMGYSABEIVGK-----HNSL-----FVPAERDGGAYR-----QFWDI
116 Mmag_103-220 -----A-FYEQIAALIRKSQA-MIEF-KPDGTVLTANDNRLQMGYSABEIVGK-----HHSI-----FVDAEVRSPQYQ-----EFWKA
117 Xcam_138-255 -----A-FEGRIDALIRKSQA-MIEF-SLDGTVLTANDNRLQMGYSABEIVGK-----HHSI-----FVPEPAQDRSSEYQ-----VFWKE
118 Xcit_161-278 -----A-FEGRIDALIRKSQA-MIEF-SLDGTVLTANDNRLQMGYSABEIVGK-----H-RM-----FVDAGTRQSEYQ-----AFWDS
119 Xcam_262-377 -----S-ADGRQALIA-VMG-MIEF-DLDGVLNANDNRLQMGYSABEIVGK-----H-SV-----FVDAAYASSEYQ-----QFWAT
120 Xcit_285-400 -----S-ADGRQALIA-VMG-MIEF-DLDGVLNANDNRLQMGYSABEIVGK-----H-V-----FVDAAYASSEYQ-----QFWAK
121 Xcam_18-133 -----S-LQKATAM-RVMA-MIEF-DLEGRILNANDNRLQMGYSABEIVGK-----HHRM-----FVHPSEERSSEYQ-----QFWDI
122 Xcit_41-156 -----S-LRHKVAAVRVMA-MIEF-DLDGVLNANDNRLQMGYSABEIVGK-----H-M-----FVTAADRSEYQ-----HFWKE
123 Atum_8-124 -----A-NACAVLAALSKSQA-MIEF-DLTGRILNANDNRLQMGYSABEIVGK-----HHSM-----FVPEPDRVSSADYQ-----AFWAK
124 Atum_13-129 -----A-NACAVLAALSKSQA-MIEF-DLSCGRILNANDNRLQMGYSABEIVGK-----HHSM-----FVPEPDRVSSADYQ-----AFWAK
125 Atum_10-124 -----S-AVALLAALSKSQA-MIEF-DLSCGRILNANDNRLQMGYSABEIVGK-----HHSI-----FVPEPDRVSSADYQ-----AFWAK
126 Atum_9-125 -----L-DASAVLDALSRQA-MIEF-DLTGRILNANDNRLQMGYSABEIVGK-----HHSI-----FVPEPDRVSSADYQ-----AFWAK
127 Sent_8-124 -----R-DARQIDALIRKSQA-MIEF-DLSCGRILNANDNRLQMGYSABEIVGK-----H-RI-----FCAPEFVATSEYQ-----EFWAR
128 Atum_8-124 -----A-DNRNMLDALIRKSQA-MIEF-DLSCGRILNANDNRLQMGYSABEIVGK-----H-RI-----FVDEIAASHAYQ-----EFWES
129 Cre_43-156 -----S-QKIDALDKS-A-MIEF-DVKGVLNANLFLQMGYSABEIVGK-----H-SI-----FVPEPDRVSSADYQ-----DFWMO
130 Ssp_21-138 -----G-EQAMIDAVSRVMA-MIEF-ELDGTIRTANENFLKVTGYRLLEIVGK-----HHRM-----FCDDPDVNSSEYQ-----AFWAK
131 Psyr_141-256 -----S-YEGKVAAIDRSQGG-MIEF-DLNGRVLTANDNRLQMGYSABEIVGK-----H-M-----FCEDDYVNSSEYQ-----AFWAK
132 Psyr_141-256 -----S-YEGKVAAIDRSQGG-MIEF-DLNGRVLTANDNRLQMGYSABEIVGK-----H-M-----FCEDDYVNSSEYQ-----AFWAK
133 Psyr_263-378 -----S-SAGKVTAIRSRQA-MIEF-DLTGKVLTANDNRLQMGYSABEIVGK-----HHRM-----FCSEEFVNSSEYQ-----ELWEK
134 Psyr_263-378 -----S-YAGKVTAIRSRQA-MIEF-DLTGKVLTANDNRLQMGYSABEIVGK-----H-RM-----FCSEEFVNSSEYQ-----ELWEK
135 Psyr_18-134 -----S-DHSIMTAIRSRQA-MIEF-DLDGVLNANDNRLQMGYSABEIVGK-----HHRM-----FCTPEHASSEYQ-----EFWEK
136 Psyr_21-134 -----S-SIMVAI-RSQA-MIEF-DLEGNILNANDNRLQMGYSABEIVGK-----H-M-----FCTPEHASSEYQ-----EFWEK
137 Naro_242-357 -----S-QSEQIAALIRSHA-MIEF-DLEGNILNANDNRLQMGYSABEIVGK-----H-RI-----FCEPELVNSSEYQ-----SLWER
138 Vcho_184-299 -----S-RSQMNANVLTQA-MIEF-TLDGTVLTANDNRLQMGYSABEIVGK-----HHSM-----FVDEQKQSOEYQ-----HFWOR
139 Cre_20-135 -----S-LEGIVAAIRSRQA-MIEF-NLDGTVLTANDNRLQMGYSABEIVGK-----HHSI-----FVDPAPAGSEYQ-----QFWOR
140 Bbac_38-159 -----S-SIREETKALHRVQA-MIEF-NLDGTVLTANDNRLQMGYSABEIVGK-----H-SM-----FCEPEYNSSEYQ-----KFQWI
141 Bbac_165-281 -----S-EYEGKVTAIRSRQA-MIEF-NLDGTVLTANDNRLQMGYSABEIVGK-----H-M-----FCDPVVNSSEYQ-----MFWEK
142 Ssp_264-381 -----R-LNAGVILNANDNRLQMGYSABEIVGK-----K-RI-----FVDETASSEYQ-----QFWQS
143 Atum_131-246 -----S-DAGKIDALIRQA-MIEF-TPTGTVLTANDNRLQMGYSABEIVGK-----H-SM-----FCEPSYTSSEYQ-----NFWKM
144 Atum_131-246 -----S-DAGKIDALIRQA-MIEF-TPTGTVLTANDNRLQMGYSABEIVGK-----H-SM-----FCDPAYTSSEYQ-----QFWOR
145 Atum_132-247 -----S-DDGKLAALIRQA-MIEF-TPDGTVLTANDNRLQMGYSABEIVGK-----H-SI-----FCEPAYTSSEYQ-----QFWKS
146 Smel_131-246 -----S-DAGKIDALIRQA-MIEF-TPDGTVLTANDNRLQMGYSABEIVGK-----H-SM-----FCEPAYTSSEYQ-----QFWKS
147 Atum_131-246 -----S-DAGKIDALIRQA-MIEF-TPDGTVLTANDNRLQMGYSABEIVGK-----H-SM-----FCDPAYTSSEYQ-----QFWKS
148 Cre_163-278 -----S-RTAKLDAVERVQA-MIEF-TVDGVLNANDNRLQMGYSABEIVGK-----HHSM-----FVDPAPAGSEYQ-----AFWAK
149 Ssp_509-626 -----Q-DENCKLEAISNASHA-MIEF-TPDGTVLTANDNRLQMGYSABEIVGK-----H-RM-----FCERDYASSEYQ-----AFWER
150 Psyr_385-500 -----S-DQGVNATIRQA-MIEF-DMAGTVLTANDNRLQMGYSABEIVGK-----H-RI-----FCEPEYNSSEYQ-----EFWKG
151 Psyr_385-500 -----S-DQGVNATIRQA-MIEF-DMAGTVLTANDNRLQMGYSABEIVGK-----H-RI-----FCEPEYNSSEYQ-----EFWKG
152 Bbac_288-403 -----S-DAGKIDALIRQA-MIEF-NMDGTVLTANDNRLQMGYSABEIVGK-----H-M-----FCNSEYTAGPYQ-----NFWAK
153 Naro_244-479 -----S-YQAKVVAIRNRLA-MIEF-DLDGVLNANDNRLQMGYSABEIVGK-----H-M-----FCAPDYVNSSEYQ-----EFWLK
154 Cvio_18-135 -----E-NOKNVLDALNQSTA-MIEF-SPDGTVLTANDNRLQMGYSABEIVGK-----H-M-----FCLPQFTASSEYQ-----AFWHD
155 Cjej_24-137 -----S-DILRSIGNTMA-MIEF-TDGVILNANDNRLQMGYSABEIVGK-----H-M-----FCLPEVNSSEYQ-----QFWKD
156 Ssp_18-132 -----S-QKIRAMTENTQA-MIEF-KVDGTVLTANDNRLQMGYSABEIVGK-----H-SI-----FVYPSFVRKDAYQ-----QFWRD
157 Mdeg_17-134 -----K-ELEKVNARSALA-MIEF-TPEGVETANDNRLQMGYSABEIVGK-----H-SI-----FVYDEYNSSEYQ-----NFWSN
158 Cre_1-103 -----S-MLEI-RPNA-VVRANAFQRLTGYSABEIVGK-----H-SI-----FLAEGTDSSEYQ-----EFWRA
159 Cjej_142-259 -----L-DLNTIAANRSMMA-MIEF-KPDGTVLTANDNRLQMGYSABEIVGK-----H-SM-----FCDSNVRSSEYQ-----QFWED
160 Agam_51-166 -----M-TKALFEALDRSLA-MIEF-EPDGTILNANDNRLQMGYSABEIVGK-----Q-RM-----FVDELFEYR--BNP-----DFWQE
161 Agam_116-231 -----L-SQATIKALERSLA-MIEF-TPNGTVLTANDNRLQMGYSABEIVGK-----H-RM-----FCDDAFYR--BNP-----RFBWE
162 Vcho_145-260 -----D-KQAAVPHSLDRSSA-MIEF-NPDGTILNANDNRLQMGYSABEIVGK-----H-RM-----FCDDAFYR--BNP-----RFBWE
163 Vpar_135-250 -----E-CKEAILNALDLSLA-MIEF-DREGTVLTANDNRLQMGYSABEIVGK-----H-KL-----FCDFDYR--BNP-----GFWKD
164 Vcho_174-289 -----E-SORDLLTAL-QNFA-MIEF-EPDGTVLTANDNRLQMGYSABEIVGK-----H-RI-----FCDFDYR--BNP-----DFWKS
165 Paer_139-254 -----L-LNAINDSIROSMA-MIEF-TPDGEILNANDNRLQMGYSABEIVGK-----H-RM-----LCDFDYR--BNP-----EFWAK
166 Paer_139-254 -----L-LNAINDSIROSMA-MIEF-TPDGEILNANDNRLQMGYSABEIVGK-----H-RM-----LCDFDYR--BNP-----EFWAK
167 Paer_20-135 -----S-ERLHMAALDRSMA-MIEF-DPDGTVLTANDNRLQMGYSABEIVGK-----H-RQ-----LCDGAYASSEYQ-----RFWER
168 Paer_3-112 -----S-ALDRSMA-MIEF-DPDGTVLTANDNRLQMGYSABEIVGK-----H-RQ-----LCDGAYASSEYQ-----RFWER
169 Cvio_142-259 -----E-DSRALRQALDRSMA-MIEF-DLDGVLNANDNRLQMGYSABEIVGK-----H-RM-----LCDEPSYAGSEYQ-----KLWQK
170 Ypes_136-253 -----Q-EHQSLLEALNRSMA-MIEF-TPQGVILNANDNRLQMGYSABEIVGK-----H-SI-----LCLEPFAHSEYQ-----QFWOR
171 Ypes_132-249 -----Q-EHQSLLEALNRSMA-MIEF-TPQGVILNANDNRLQMGYSABEIVGK-----H-SI-----LCLEPFAHSEYQ-----QFWOR
172 Psyr_20-135 -----S-LKGLTSALEKSMA-MIEF-GLDGKILNANDNRLQMGYSABEIVGK-----T-D-----FCEPEVLSSEYQ-----DLWAS
173 Psyr_20-135 -----S-LKGLTSALEKSMA-MIEF-GLDGKILNANDNRLQMGYSABEIVGK-----T-D-----FCEPEVLSSEYQ-----DLWAS
174 Pflu_18-135 -----N-QCAGLLEALNRSMA-MIEF-DVDGTVLTANDNRLQMGYSABEIVGK-----H-H-----FCSPEFARQNYT-----ELWKS
175 Cvio_20-137 -----R-QSRSLAALIRSMA-MIEF-SVDGTVLTANDNRLQMGYSABEIVGK-----H-H-----FQDDYANSSEYQ-----DFWQK
176 Cvio_140-257 -----H-EARNINKAVERSMA-MIEF-TPEGTVLTANDNRLQMGYSABEIVGK-----H-RI-----FPADEFVNSSEYQ-----HMQQK
177 Lmon_8-124 -----E-DATLILLOGLQNVMA-MIEF-DTNKVTYANALFAAMGYTSEEMVQL-----SHPD-----LCFDFVQTSASYR-----EMMTN
178 Lmon_8-124 -----E-DATLILLOGLQNVMA-MIEF-DTNKVTYANALFAAMGYTSEEMVQL-----SHPD-----LCFDFVQTSASYR-----EMMTN
179 Linn_8-124 -----E-DATLILLOGLQNVMA-MIEF-DTNKVTYANALFAAMGYTSEEMVQL-----SHPD-----LCFDFVQTSASYR-----EMMTN
180 Ssp_14-130 -----S-DTQVTRAIQNLMA-MIEF-DDRRVAYVNDNFARTMGYSABEIVGK-----YHRD-----FPGFADSEYQ-----LFWRK
181 Oihe_9-126 -----K-MDNLVVQALSNLA-MIEF-DLTKVAVVNDNFARTMGYSABEIVGK-----H-H-----FCEFDTFVSEYQ-----LFWNR
182 Ssp_11-127 -----L-ASADILEALIRNMA-MIEF-DTQRRVAVVNDNFARTMGYSABEIVGK-----H-H-----FCRADFANSEYQ-----KLWKS

Figure C.1 continued

Generic PAS

183	Psyr_20-135	-----LEQAGQILNLETVAVVLDG--TG--TQTVMNLFETEMSYA-----	QA	A	I	V--GRSL--SE--LSPPELS-----	GDVHQKRALTA
184	Psyr_20-135	-----LEQAGQILNLETVAVVLDG--TG--TQTVMNLFETEMSYA-----	QA	A	I	V--GRSL--SE--LSPPELS-----	GDVHQKRALTA
185	Psyr_35-150	-----LEQAGHILNLETVAVVLDG--QKQVQTVNGLFETELSYA-----	QA	A	I	A--GRSL--SE--MSPPELS-----	GDVHQKRALTA
186	Pput_26-143	M--TEQVKSLSLDEMLVLQLDP--QGRITEMVN--NFESEMLYK-----	AE	C	I	L--GRNI--ED--IVPAHVK-----	SLDFFYQRMKSA
187	Pflu_18-135	SS--LOVKESLSEMLVLTLDP--DGRITQSVN--NFLSEMLYK-----	SH	I	I	I--GRAI--ED--IVPAHVK-----	SDFFHHRFKAA
188	Vpar_19-136	YS--LOQIQSLSDEMLRLTLDA--KGNVTSAN--KFLQOOLSLS-----	DK	I	L	I--AKHI--SD--LVPSNVR-----	NTDHFYRMKQSA
189	Psyr_22-139	SMYRCOMQGMDAQMVCLTLDA--SYHIVHANDLNLSTLGYLS-----	LE	C	V	L--GKDL--DH--LVPTYVK-----	QLDCYRSRLKVA
190	Psyr_47-164	SMYRCOMQGMARMVSLSLDA--SNRIAHAN--NFLRALGYT-----	AE	C	I	L--GREL--DQ--MVPTYVK-----	QLDCYRNRLKTA
191	Psyr_17-135	-----LHVLCQLLGRIPQDMLTVQVVG--NF--TISAANQGFAPALGYT-----	PD	I	S	I--S--GFPL--SS--IAAFDSK-----	GMPWFHGLKTT
192	Vcho_63-177	QVNQAFVQVIRHIALLECEP--NGTICYAS--AFAPALCRVS-----	AE	A	M	V--GADF--AN--LWRTHQQ-----	P--SVQRLLQD
193	Ssp_389-504	KTLSLVANETD--NSVLIADA--DGRIEYVN--GFKTLTGHE-----	YK	I	V	I--GKKP--GE--LQGRHT-----	DPETKRIKIREN
194	Psyr_89-209	-----RASSEGGLWDMDEVVAGDPVN--PNNRFWWSQCFRTLLGFN-----	DE	R	D	P--NVLASWADRLHPQDKQ-----	ASL--DAFAKH
195	Psyr_88-209	-----NRASSEGGLWDMDEVVAGDPVN--PNNRFWWSQCFRTLLGFN-----	DE	R	D	P--NVLASWADRLHPQDKQ-----	ATL--DAFAKH
196	Cvio_92-213	-----NRASSEGGLWDMDEVVAGDPVN--PNNRFWWSQCFRTLLGFN-----	DE	R	D	P--NVLASWADRLHPQDKQ-----	ATL--DAFAKH
197	Psyr_231-348	-----SDGLWDMDEVVAGDPVN--ARNPFWWSQCFRTLLGFE-----	TV	E	E	P--DVLDSWASRLHPDEKE-----	ASL--TAFGAH
198	Psyr_227-348	-----REMLSDGLWDMDEVVAGDPVN--ARNPFWWSQCFRTLLGFE-----	TV	E	E	P--DVLDSWASRLHPDEKE-----	ASL--TAFGAH
199	Cvio_231-352	-----SEMLAEGGLWDMDEVVAGDPVN--GAHAFWWSQCFRTLLGFD-----	SE	C	E	P--DVLDSWASRLHPDEKE-----	SVL--DAFAKH
200	Rrub_41-162	-----DGNAGVGLWDVAVLHGG--PLH--AQSRWTSAAEFRLVGFK-----	SE	T	E	P--NVVGSWADRLHPEDAA-----	STE--DAFAGAL
201	Rrub_173-294	-----SHHAGVGLWDVAVLHGG--AMH--POSRTWTSAAEFRLVGFK-----	SE	A	E	P--DVLDSWASRLHPDEKE-----	PTT--AAAFGLN
202	Rpal_23-144	-----DTL--LHCGIGLWDAILHGG--AMH--PKARWTSAAEFRLVGFK-----	SE	A	E	P--NVVGSWADRLHPEDAA-----	PTT--AAAFGLN
203	Rrub_28-149	-----TRAGVGLWDVKHGSDDLH--PLSVWWSSEELRLGFT-----	DA	A	E	P--DVLDSWASRLHPDEKE-----	DAL--DAFAGAL
204	Mmaz_29-147	-----KEISLLID--LPVTVFRIGNES--SWAHIKCKSVQOLGYK-----	KN	I	F	I--TR--TWSLCLCPEDAL-----	ALN--FVYOKA
205	Mace_31-149	-----KEVSSLLID--LPVTVFRIGNES--SWAHIKCKSVQOLGYK-----	KN	I	F	I--TR--TWSLCLCPEDAL-----	ALN--FVYOKA
206	Mmaz_29-147	-----L--TGWLLID--LPVTVFRIGNES--SWAHIKCKSVQOLGYK-----	KN	I	F	I--TR--TWSLCLCPEDAL-----	ALN--FVYOKA
207	Mbur_32-150	-----EISLLID--LPVTVFRIGNES--SWAHIKCKSVQOLGYK-----	KN	I	F	I--TR--TWSLCLCPEDAL-----	ALN--FVYOKA
208	Mmaz_153-269	-----ESQKAIVSIPKPSLALVDA--SGKIKYINDYFVVKQKFKS-----	AS	I	A	I--GLSP--AD--LMES-----	NNK--KSIABTV
209	Mace_155-271	-----ESQRAIVSS--PQPSLALVDA--SGKIKYINDYFVVKQKFKS-----	AE	D	A	I--GVSP--TD--H--ES-----	GDK--KSIABTV
210	Mmaz_153-269	-----ESQRAIVS--IPRPSLALVDA--SGKIKYINDYFVVKQKFKS-----	AE	A	A	T--GRSP--SE--LLDT-----	SSK--KSIABTV
211	Mbur_156-284	-----ESQKIVKSIPEPALAIVDP--EGKVKHINDYFVVKQKFKS-----	DE	D	V	I--GRSP--SE--LIGK--TLVGDNSNM--LGAS--TIVSABV-----	EGG--TIVADKV
212	Aful_169-283	-----ELVKEVFNKMPFPAVYVLFVR--DHKIQYANDYFVVKQKFKS-----	AE	A	I	I--GLSP--SE--LFTK-----	EGG--TIVADKV
213	Aful_189-305	-----ELIKEVFN--PT--YVIFVGE--DGLIKFAN--NVARLAGFET-----	AE	E	V	V--GLRP--AL--VAVIHKD-----	Y--LDNK--KRVN--BN
214	Aful_62-177	-----KFVRELVRQIPKPAFVFLNKK--DGILIEYNEVYAAEVGAE-----	IS	C	M	I--GRKP--SE--LASNV-----	AAGK--KTFVELA
215	Aful_35-157	-----AFIEALIR--VPKH--VEFYFLDA--EGKRLVY--DYNLFBFYGS-----	RE	C	V	I--GKRP--SE--LFLDPTG--QKTL--EVGM--KMTIETA-----	RTD--KACACRA
216	Hsp_8-123	-----GALSQFFATPEP--L--FVVR--DG--VVLWMDAIVMTGVE-----	AA	I	A	V--GTPS--TE--LFTGTE-----	KET--QTILASAV
217	Aful_319-440	-----TPV--AMVYID--NHN--VYVWNAAEELTGK-----	AE	E	M	V--GTRK--TW--YFYPD-----	QR--PILADLV
218	Ddes_384-499	-----QRYVDVINTVDPD--IFAV--E--EYNIIMANKATQDFGLG-----	IN	K	K	C--DC--C--HD--QFSTVC-----	QTE--KCPIDMA
219	Dul_426-541	-----EGYKNIVNAIPDP--VFAVDD--DYNILLANNAVARLAGT-----	IGBQ	V	R	G--M--C--SS--IFNTSVC-----	GTD--PICQA
220	Dpsy_501-615	-----DEKIENTN--IPTP--ILSITD--DYNITFMP--PAGAAGVGT-----	PD	E	V	I--GKRC--YD--LFTKPHC-----	RTD--KACACRA
221	Dpsy_379-493	-----N--KIENLN--IPTP--IMSITD--EY--VTFMPPAAAAGVGT-----	PD	E	V	I--GKRC--YD--LFTKPHC-----	RTD--KACACRA
222	Dpsy_257-371	-----N--KIENLN--IPTP--IMAITD--DYNITFMP--PAGAAGVGT-----	PD	E	V	I--GKRC--YD--LFTKPHC-----	RTD--KACACRA
223	Dpsy_135-249	-----YQMOTDLNVIPTP--IMELIK--SYNITFMP--PAGAAGVGT-----	PD	E	V	I--GKRC--YD--LFTKPHC-----	RTD--KACACRA
224	Dpsy_623-736	-----DEKIENTN--IPTP--ILSITD--DYNITFMP--PAGAAGVGT-----	PD	E	V	I--GKRC--YD--LFTKPHC-----	RTD--KACACRA
225	Dvul_393-508	-----GFMQGLIRGIONP--FAVVD--TL--IRNCOSQMAVIT--STGAD-----	MA	D	V	--GMIH--SE--FLPADRN-----	R--ALLHDV
226	Dvul_261-373	-----CYNRSVLEGLIIV--LAVV--A--NRRIEFIN--PACTIV--STGAD-----	Y	S	S	K--GKDF--SA--FMQCGGV-----	R--EDITATV
227	Ddes_360-475	-----AFSRSVLEGLIIV--LAVV--A--NRRIEFIN--PACTIV--STGAD-----	Y	S	S	K--GKDF--SA--FMQCGGV-----	R--EDITATV
228	Dvul_359-474	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
229	Dvul_118-232	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
230	Dvul_118-232	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
231	Dvul_384-499	-----GLTQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
232	Ddes_263-378	-----GLTQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
233	Dvul_366-480	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
234	Dvul_389-504	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
235	Ddes_379-494	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
236	Dul_405-518	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
237	Xcam_140-257	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
238	Xcit_156-271	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
239	Xcam_163-285	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
240	Xcit_164-284	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
241	Xcam_145-249	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
242	Xcam_277-395	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
243	Neur_277-373	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
244	Vvul_42-160	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
245	Vvul_38-156	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
246	Vcho_38-157	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
247	Vvul_169-288	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
248	Vvul_165-284	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
249	Sep_146-263	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
250	Vvul_296-413	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
251	Vcho_164-282	-----GFAQGVNLGIVV--VVLVD--DEKATIN--PCLMDI--QDRTG-----	PE	C	C	L--GRSL--AD--LFDVDPG-----	R--KTAVGRS
252	Rrub_9-120	-----E--FREAL--QLPVP--VILAR--DMVISWMNRATRSERLRI--LHLP-----	VAPE	I	V	V--GQSI--DV--FHRNPA-----	R--KAI
253	Rrub_130-241	-----RN--MLD--LPVN--VMLADP--KTMILTYANRSLDTLAKLQHLS-----	FVSQ	I	V	--GRSI--DI--FHRN--SH-----	R--KAI
254	Mdeg_261-371	-----SOMCQMTAR--VMMASG--PDLTISY--NKASRELLSTLSEH--P-----	CPAA	E	V	V--GKSV--DI--FHRNPA-----	R--KAI
255	Mdeg_138-254	-----S--QASALK--LCOAN--VMLADN--ELNIVVMNDTVMVLMQEREKELOQ--HL--PSFKV-----	Y	I	I	--GTNI--DI--FHRN--SH-----	R--KAI
256	Rpal_13-128	-----L--TLATKAVRAN--IMVSDP--ELDIVVMNDTVMVLMQEREKELOQ--HL--PSFKV-----	Y	I	I	--GTNI--DI--FHRN--SH-----	R--KAI
257	Sone_145-261	-----QRLLEALNNTSTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
258	Mdeg_527-640	-----IK--L--DVTSTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
259	Mdeg_269-384	-----ARVRYALDVTAN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
260	Cvio_457-573	-----ARISALD--CTTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
261	Paer_173-289	-----ARISALD--CTTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
262	Paer_173-289	-----ARISALD--CTTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
263	Daro_436-552	-----LRKIGL--NVTAN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
264	Sone_21-137	-----QRYFOLLN--NRNN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
265	Daro_261-377	-----TRIKVALD--VSSN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
266	Cvio_181-297	-----ARISALD--CTTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
267	Cvio_319-435	-----ARISALD--CTTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
268	Cvio_43-159	-----LRVRNALD--CTTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
269	Lint_219-335	-----LRVRNALD--CTTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
270	Lint_219-335	-----LRVRNALD--CTTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
271	Lint_349-466	-----LRVRNALD--CTTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
272	Lint_481-597	-----LRVRNALD--CTTN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
273	Cvio_19-144	-----ROLESVL--HADN--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI
274	Gul_12-134	-----LDVLKQML--EVKNI--VMIADN--NRTIYVMNRSVEMALRRSESEILQ--VL--PHFSV-----	X	I	L	--GSSM--DI--FHRN--SH-----	R--KAI

PAS_Aer

Figure C.1 continued

1	Daro_8-120	LKA	-----G-----	RPWRGVVKNRR	KDGG	YYWVLNANSP	IEH-G--QI	VGYSVRLTPGR	-----	46140309
2	Daro_4-120	LKA	-----G-----	RPWRGVVKNRR	KDGG	FYVVVANASPI	IEH-G--QV	VGYSVRLTPGR	-----	53729639
3	Reut_6-119	LKA	-----G-----	TPWTGVVKNRR	KDGG	FYWVLNANSP	IEH-G--QV	SGYLSVRLTPKA	-----	53762384
4	Rgei_11-124	LKA	-----G-----	RPWTGLVKNRR	SSGG	AFWVKANI	IPMFKD-R--QT	VGFTSVQCPADA	-----	47573623
5	Bcep_8-121	LKA	-----G-----	EPWTALVKNRR	RKNGD	HYWVRANATPV	VMRN-G--QB	QGYMSVRLTPKASRD	-----	46322895
6	Bcep_8-121	LKA	-----G-----	EPWTALVKNRR	RKNGD	HYWVRANATPV	VMRN-G--QB	QGYMSVRLTPKATRD	-----	46327934
7	Bfun_8-121	LKA	-----G-----	EPWSALVKNRR	RKNGD	HYWVRANATPV	VMRN-G--QB	QGYMSVRLTPQASRE	-----	48783787
8	Rsol_8-121	LKA	-----G-----	EPWSALVKNRR	RKNGD	HYWVRANATPV	VMRN-G--QB	QGYMSVRLTPKPTRD	-----	17431698
9	Bcep_8-121	LKA	-----G-----	EPWTALVKNRR	RKNGD	HYWVRANATPV	VMRN-G--QB	QGYMSVRLTPKAPHD	-----	46317933
10	Bcep_6-121	LKA	-----G-----	EPWTALVKNRR	RKNGD	HYWVRANATPV	VMRN-G--QB	QGYMSVRLTPKAPRD	-----	46322894
11	Bmal_1-101	LKA	-----G-----	EPWTALVKNRR	RKNGD	HYWVRANATPV	VMRN-G--QB	QGYMSVRLTPKAPRD	-----	52423246
12	Bpse_8-121	LKA	-----G-----	EPWTALVKNRR	RKNGD	HYWVRANATPV	VMRN-G--QB	QGYMSVRLTPKAPRD	-----	52211725
13	Bmal_8-121	LKA	-----G-----	RSWTAVIKNRR	KDGG	HYWVRANATPV	VMRN-G--QB	QGYMSVRLTPKPSRE	-----	52428136
14	Ecol_8-121	LKA	-----G-----	EPWSGIVKNRR	RKNGD	HYWVRANATPV	VMRE-G--KT	SGYMSIRTRATDE	-----	26110084
15	Ecol_8-121	LKA	-----G-----	EPWSGIVKNRR	RKNGD	HYWVRANATPV	VMRE-G--KT	SGYMSIRTRATDE	-----	13363427
16	Styp_8-121	LKA	-----G-----	EPWSGIVKNRR	RKNGD	HYWVRANATPV	VMRE-G--RV	SGYMSIRTRATDD	-----	16421774
17	Styp_8-121	LKA	-----G-----	EPWSGIVKNRR	RKNGD	HYWVRANATPV	VMRE-G--RV	SGYMSIRTRATDD	-----	16504293
18	Ecar_8-120	LKA	-----G-----	DSWTGLVKNRR	RKNGD	HYWVRANATPV	YQO-E--QL	AGYLSVRLTPNA	-----	49613026
19	Ypes_8-120	LKA	-----G-----	LSWSSIVKNRR	RKNGD	HYWVRANATPV	LRN-G--RL	SGYLSVRLTPATR	-----	15979682
20	Ecar_8-121	LKA	-----G-----	KIWTAVVKNRR	KSGG	YYWVKASTT	PLMKR-G--KT	SGYMSVRLTPVQSE	-----	49611454
21	Ecar_8-118	LKA	-----G-----	NIWTGLVKNRR	RKNGD	HYWVKSSSTT	PLRKG-G--ET	SGYMSVRLTPA	-----	49611455
22	Reut_8-121	LKA	-----G-----	RSWVGIVKNRR	RKNGD	YYWVSATV	TPTHID-G--RL	VGYSVRLTPMATRE	-----	53762126
23	Rmet_8-121	LKA	-----G-----	KSWVGIVKNRR	RKNGD	HYWVQATV	TPTRVG-D--RV	VGYSVRLTPMASRE	-----	48769261
24	Avin_10-121	LKA	-----G-----	LGWNGIVKNRR	RKNGD	HYWVRANATPV	YEG-D--RL	VGYSVRLTPRASRR	-----	23210254
25	Rsol_8-121	LKA	-----G-----	QSWGIVKNRR	RKNGD	HYWVQANV	TPVLQD-G--AT	AGYTSVRLTPATPA	-----	17430723
26	Bbro_8-121	LKA	-----G-----	QSWLGIVKNRR	RKNGD	HYWVLNANATPV	YED-G--EV	VAYSSVRLTPKASDE	-----	33755988
27	Tden_8-121	LKA	-----G-----	LPWTGLVKNRR	RKNGD	HYWVQANV	TPVRED-G--VI	SGFLSVRLTPRE	-----	52006288
28	Bjap_8-121	LKA	-----G-----	KPWLGAIVKNRR	RKNGD	FYWLATASATP	IREN-G--IV	SGYTSIRTRLPAD	-----	27351190
29	Rpal_8-121	LKA	-----G-----	RPWGAIVKNRR	RKNGD	FYWLASATP	IREN-G--QV	SGYMSVRLTPKPAD	-----	39648595
30	Daro_8-121	LKA	-----G-----	RPWTGMVKNRR	RKNGD	YYWVLATV	TPLEGG-G--ET	LGYSVRLTPKASQA	-----	53759254
31	Neur_4-121	LKA	-----G-----	RPWSGMVKNRR	KDGG	CYWVYANV	TPIREH-G--IV	TGHSVRLTPKPTRD	-----	30180852
32	Mfia_8-121	LKA	-----G-----	LPWTGMVKNRR	RKNGD	YYWVLNANATPV	IREN-G--NV	VGYSVRLTPKPSRQ	-----	53760232
33	Tden_8-121	LKA	-----G-----	KVWSGIVKNRR	RKNGD	YYWVDADV	APLIEN-G--KT	VGYSIRTRATPRE	-----	52007875
34	Rgei_8-120	LKA	-----G-----	QPSWGVKNRR	RKNGD	CYWVIANV	TPIMSG-D--QB	SGYMSVRLTPADR	-----	47574282
35	Bfun_4-120	LKA	-----G-----	RPWTALVKNRR	RKNGD	HYWVHANV	TPVVEKG-G--TV	VGYSVRLTPKPER	-----	48826263
36	Bbro_8-121	LKA	-----R-----	KPWLGAIVKNRR	KSGG	FYWLANAM	MEVIEA-G--NV	SGYASVRLTPKATQA	-----	33576529
37	Lint_8-121	LKA	-----G-----	NPWSGILIKNRR	KSGG	YYWVDATV	TPVMNE-G--VI	SGYMSVRLTPKATED	-----	24193388
38	Psyr_8-121	LKA	-----G-----	QPMGIVKNRR	RKNGD	HYWVNAYV	TPVLEN-R--QV	VGFSVRLTPKPTAE	-----	23471893
39	Psyr_8-121	LKA	-----G-----	QPMGIVKNRR	RKNGD	HYWVNAYV	TPVLEN-R--QV	VGFSVRLTPKPTAE	-----	28852458
40	Pput_8-121	LKA	-----G-----	QPMGIVKNRR	RKNGD	HYWVNAYV	TPFDN-N--QV	VGFSVRLTPKPTAE	-----	24983638
41	Pflu_8-121	LKA	-----G-----	LPWMGIVKNRR	KDGG	HYWVNAYV	TPVFEQ-N--QV	VGYSVRLTPKPTAE	-----	48731902
42	Paer_8-121	LKA	-----G-----	RPWGMVKNRR	RKNGD	HYWVSAYV	TPYDQ-G--IV	VGYSVRLTPKPTAE	-----	53727763
43	Psyr_8-122	LKA	-----G-----	KPWMGVKNRR	RKNGD	HYWVSAYV	TPVAYEN-G--RL	VGYSVRLTPKPTRDQ	-----	46187912
44	Psyr_8-125	LKA	-----G-----	KPWMGVKNRR	RKNGD	HYWVSAYV	TPVAYEN-G--RL	VGYSVRLTPKPTRDQ	-----	28852093
45	Pput_8-121	LKA	-----G-----	KPWMGIVKNRR	RKNGD	HYWVSAYV	TPVAYEQ-G--RL	SGYSVRLTPKPTR	-----	24986255
46	Pput_8-120	LKA	-----G-----	KSWMGIVKNRR	RKNGD	HYWVNAYV	TPILEG-G--RV	VGYSVRLTPKPTR	-----	4545127
47	Pput_8-120	LKA	-----G-----	KSWMGIVKNRR	RKNGD	HYWVNAYV	TPILEG-G--RV	VGYSVRLTPKPTR	-----	24983779
48	Vvul_8-121	LKA	-----G-----	KSWMGIVKNRR	RKNGD	HYWVDFAF	SPIDKSSG-KV	VEYQSVRLTPCSR	-----	37200780
49	Vcho_8-120	LKA	-----G-----	KSWMGIVKNRR	RKNGD	HYWVDFAF	SPIDKSSG-KV	VEYQSVRLTPCSR	-----	9658429
50	Vpar_4-117	LKA	-----G-----	KHWMGMVKNRR	RKNGD	HYWVDFAF	SPIDKSSG-KV	VEYQSVRLTPCSR	-----	28808762
51	Ppro_1-115	LKA	-----G-----	KSWMGLVKNRR	RKNGD	HYWVSAYV	TPITNDKG-EV	VEYQSVRLTPCSR	-----	46916433
52	Vvul_5-119	LKA	-----G-----	KSWMGLVKNRR	RKNGD	HYWVSAYV	TPITNDKG-EV	VEYQSVRLTPCSR	-----	37201131
53	Sone_6-121	LKA	-----G-----	EPWKGIVKNRR	RKNGD	HYWVDAYV	TPSPIMIN-G--QV	VEYQSVRLTPCSR	-----	24346080
54	Sthe_8-121	LKA	-----G-----	ERWVGIVKNRR	RKNGD	HYWVKAFV	SPVLED-G--KT	IGYRSVRLTPKPTR	-----	51856248
55	Sone_14-127	LKA	-----G-----	QSWGIVKNRR	RKNGD	HYWVDAYV	TPSPIMIN-G--KT	IGYRSVRLTPKPTR	-----	24347109
56	Ecol_6-121	LKA	-----G-----	RSWGLVKNRR	RKNGD	HYWVSAYV	TPVIAKN-G--SI	VEYQSVRLTPKPEPE	-----	48195
57	Vcho_11-124	LKA	-----G-----	RSWGLVKNRR	RKNGD	HYWVSAYV	TPVIAKN-G--SI	VEYQSVRLTPKPEPE	-----	9654939
58	Pflu_6-123	LKA	-----G-----	RSWGLVKNRR	RKNGD	HYWVSAYV	TPVIAKN-G--SI	VEYQSVRLTPKPEPE	-----	48730183
59	Vpar_8-121	LKA	-----G-----	KAWRGVKNRR	RKNGD	HYWVDAYV	TPVIAKN-G--SI	VEYQSVRLTPKPEPE	-----	28809550
60	Vvul_8-121	LKA	-----G-----	KAWRGVKNRR	RKNGD	HYWVDAYV	TPVIAKN-G--SI	VEYQSVRLTPKPEPE	-----	37201308
61	Vcho_24-136	LKA	-----G-----	HAWRGVKNRR	RKNGD	HYWVDAYV	TPVIAKN-G--SI	VEYQSVRLTPKPEPE	-----	9658074
62	Ppro_6-119	LKA	-----G-----	KAWRGVKNRR	RKNGD	HYWVDAYV	TPVIAKN-G--SI	VEYQSVRLTPKPEPE	-----	49615609
63	Ppro_8-121	LKA	-----G-----	KAWRGVKNRR	RKNGD	HYWVDAYV	TPVIAKN-G--SI	VEYQSVRLTPKPEPE	-----	28809550
64	Sone_6-119	LKA	-----G-----	KAWRGVKNRR	RKNGD	HYWVDAYV	TPVIAKN-G--SI	VEYQSVRLTPKPEPE	-----	37201308
65	Rpal_7-121	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	39937542
66	Rpal_6-120	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	39937544
67	Rpal_7-121	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	39937263
68	MMag_1-109	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	46200832
69	Bjap_7-122	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	27378086
70	Rpal_1-124	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	39937372
71	MMag_20-133	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	46203509
72	MMag_8-122	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	23015159
73	Wsuc_9-123	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	34557986
74	Wsuc_7-121	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	34557696
75	Wsuc_7-121	LKA	-----G-----	REIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	34557107
76	Daro_1-109	LKA	-----G-----	KEIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	46140878
77	Cjei_1-113	LKA	-----G-----	KEIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	1592513
78	Cvio_9-125	LKA	-----G-----	HEIFGVKNRR	RKNGD	HYWVFAHVT	TPSPFDHGG-NI	VGYHSNRRTPDPR	-----	34495851

PAS_Che

79	Ypes_15-131	LKA	-----G-----	RPISDNIKRR	KDGG	VINLQGTYP	VPVDRQG-N--N	VIEIIKIASDVTER	-----	21958470
80	Ypes_1-112	LKA	-----G-----	RPISDNIKRR	KDGG	VINLQGTYP	VPVDRQG-N--N	VIEIIKIASDVTER	-----	15980509
81	Ypes_11-127	LKA	-----G-----	RPISDNIKRR	KDGG	VINLQGTYP	VPVDRQG-N--N	VIEIIKIASDVTER	-----	51590152
82	Paer_17-134	LKA	-----G-----	TPQGRVFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	53727981
83	Paer_17-134	LKA	-----G-----	TPQGRVFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	9479125
84	Vpar_13-130	LKA	-----G-----	KHKRGVFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	28808720
85	Vcho_53-169	LKA	-----G-----	KSHSGVFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	9658296
86	Vcho_23-140	LKA	-----G-----	QAQKGMFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	9655902
87	Pput_23-138	LKA	-----G-----	KAISGVFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	24985060
88	Pflu_23-138	LKA	-----G-----	EPISGVFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	48730737
89	Agam_1-111	LKA	-----G-----	RVKSGVFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	31195287
90	Pput_148-265	LKA	-----G-----	EYIAERFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	24982183
91	Pflu_140-257	LKA	-----G-----	EFVAGRFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	48730078
92	Psyr_145-261	LKA	-----G-----	EFVAGRFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	46188225
93	Psyr_170-286	LKA	-----G-----	EFVAGRFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	28851464
94	Psyr_140-257	LKA	-----G-----	EFVAGRFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	46188835
95	Psyr_155-272	LKA	-----G-----	EFVAGRFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	28853310
96	Vpar_141-258	LKA	-----G-----	QFVDRFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	28808847
97	Psyr_125-242	LKA	-----G-----	VFVAGRFKRLR	RKNGD	PIWLEATY	FPVKNAEG-A--A	VVEVLKIAADVTRN	-----	28854157

Figure C.1 continued

98	Psyr_140-257	LNQ	-----G-----	EYVEGRFRM	SRG	EINLQATYNPVHDSAG--R	LYKVVKFASV	TRQ-----	28855155
99	Pput_143-260	LNK	-----G-----	EYHSHRFERV	KG	TVFLEASYNPIFDSKG--R	LCKVVKFASDIT	THQ-----	24985060
100	Pflu_143-260	LNK	-----G-----	EYHSHRFERK	KG	TVYLEASYNPLFDAKG--R	LYKVVKFASDIT	THQ-----	48730737
101	Psyr_144-257	LNK	-----G-----	EFIDGQFKRI	KG	GVWLEATYNPVFDDVG--K	LYKIVKFASDIT	QR-----	23471312
102	Psyr_143-257	LNK	-----G-----	EFIDGQFKRI	KG	GVWLEATYNPVFDDVG--K	LYKIVKFASDIT	AR-----	28855718
103	Paer_140-257	LNK	-----G-----	EYVVGQFRRV	HRNG	PVWLEASYNPVYDADG--K	LFKVVKFASDVS	DR-----	53727683
104	Paer_117-234	LNK	-----G-----	EYVVGQFRRV	HRNG	PVWLEASYNPVYDADG--K	LYKVVKFASDVS	DR-----	9947371
105	Pflu_144-257	LNK	-----G-----	ELFQGGFERV	KG	TVWLEANYNPVYDAAG--R	LCKVVKFASDV	ITAR-----	48730703
106	Bjap_134-256	LNK	-----G-----	EYQAGEFKRI	KG	GVWILASYNPLLDENG--K	PFVAKFATDIT	AT-----	27349173
107	Rpal_135-250	LNK	-----G-----	EYFPGEFKRI	KG	GVWILASYNPLLDARG--K	PFKVVKYATDV	ITAQ-----	39647354
108	Bjap_263-378	LNK	-----G-----	EYQAAEYKRI	KG	GVVYIQASYNPILDLNG--K	PFKVVKYATDIT	TKQ-----	27349173
109	Rpal_256-371	LNK	-----G-----	EYQAAEYKRI	KG	GVVYIQASYNPILDLNG--R	PFKVVKFATDIT	TKQ-----	39648804
110	Rpal_255-372	LNK	-----G-----	EYQAGEYKRI	KG	GVVWIIQASYNPILDLNG--R	PFKVVKYAADIT	TAQ-----	39647354
111	Rpal_133-249	LNK	-----G-----	EYQAAEYKRI	KG	GVVWIIQASYNPIFDDKG--R	PAKVVKFATDV	ITEQ-----	39648804
112	Bsp_90-205	LNK	-----G-----	EYQSAQYKRI	KG	GVVWIIQASYNPILDLNG--K	PFKVVKFATDIS	ISAQ-----	18033717
113	Rpal_11-128	LNK	-----G-----	EYQAGEFHRI	KG	GVVWIIQASYNPILDKNG--K	PTGVVVFPAADIT	ATA-----	39647354
114	Bjap_19-134	LNK	-----G-----	EYQAAEFKRI	KG	GVVWLEASYNPVPDNAG--K	PFKVVVKIATDIT	ATA-----	27349173
115	Rrub_90-205	LNK	-----G-----	EYQARQFMRI	KG	GVVWLEASYNPILNLDG--K	PFKVVKFATDIS	ISGR-----	48764139
116	Mnag_103-220	LAD	-----G-----	EYQSGQFRRV	HRNG	GVWIIQASYNPILFDSG--K	LYAVVVFANDV	TEA-----	46203332
117	Xcam_138-255	LNK	-----G-----	EYFDAGYKRV	GRDGR	GVWIIQASYNPVLDERG--R	PFKVVKYATDIT	ITRQ-----	21112820
118	Xcit_161-278	LNK	-----G-----	EYFDAGYKRI	KG	GVVWIIQASYNPVLDEHG--R	PFKVVKYATDIT	TRQ-----	21117943
119	Xcam_262-377	LNK	-----G-----	EYFDAGYKRI	KG	GVVWIIQASYNPILFDSG--R	PFKVVKYATDIT	TDQ-----	21112820
120	Xcit_285-400	LNK	-----G-----	EYFDAGYKRI	KG	GVVWIIQASYNPILFDSG--R	PFKVVKYATDIT	TDQ-----	21117943
121	Xcam_18-133	LNK	-----G-----	EYFHAGYKRI	KG	GVVWIIQASYNPILDRSG--K	PFKVVKYATDIT	ATA-----	21112820
122	Xcit_41-156	LNK	-----G-----	EYFHAGYKRI	KG	GVVWIIQASYNPILDRSG--K	PFKVVKYATDIT	ATA-----	21117943
123	Atum_8-124	LNK	-----G-----	EYFDQGYKRI	KG	GVVWLEASYNPVMR--RG--K	PFKVVVKIATDIT	ATA-----	6498287
124	Atum_13-129	LNK	-----G-----	EYFDQGYKRI	KG	GVVWLEASYNPVMR--RG--K	PFKVVVKIATDIT	ATA-----	15163624
125	Atum_10-124	LNK	-----G-----	EYFDQGYKRI	KG	GVVWLEASYNPVMR--RG--R	PFKVVVKIATDIT	TER-----	15160314
126	Atum_9-125	LNK	-----G-----	EYDQGYKRI	KG	GVVWLEASYNPVMR--FG--K	PFKVVVKIATDIT	TVI-----	15160002
127	Sent_8-124	LNK	-----G-----	EYDQGYKRI	KG	GVVWLEASYNPVMR--NG--K	PFKVVKFATDIT	ATA-----	11545452
128	Atum_8-124	LNK	-----G-----	EYDQGYKRI	KG	GVVWLEASYNPVMR--SG--K	PFKVVVKIATDIT	ATA-----	17741118
129	Core_43-156	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQASYNPVKNSAG--K	PFKVVVKIATDIT	ATA-----	13425051
130	Sep_21-138	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQASYNPILDAAG--K	PFKVVVKIATDIT	ATA-----	52010400
131	Psyr_141-256	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQASYNPILFDDG--R	PFKVVVFANDV	ITES-----	23468713
132	Psyr_141-256	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQASYNPILFDDG--R	PFKVVVFANDV	ITES-----	28852878
133	Psyr_263-378	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--K	PFKVVVFALDV	IVA-----	23468713
134	Psyr_263-378	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--K	PFKVVVFALDV	ITEA-----	28852878
135	Psyr_18-134	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--N	PFKVVVFATDV	ITAQ-----	23468713
136	Psyr_21-134	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--N	PFKVVVFATDV	ITAQ-----	28852878
137	Naro_242-357	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--K	PFKVVVFATDV	ITAQ-----	48847895
138	Vcho_184-299	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--K	PFKVVVFATDV	ITAQ-----	9634496
139	Core_20-135	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--K	PFKVVVFATDV	ITAQ-----	13423061
140	Bbac_38-159	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--N	PFKVVVFATDV	ITAV-----	39576448
141	Bbac_165-281	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--K	PFKVVVFATDV	ITAA-----	39576448
142	Sep_264-381	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--D	PFKVVVFATDV	ITDQ-----	52010400
143	Atum_131-246	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--R	PFKVVVFATDV	ITR-----	6498287
144	Atum_131-246	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--K	PFKVVVFATDV	ITGR-----	15160314
145	Atum_132-247	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--R	PFKVVVFATDV	ITDR-----	15160002
146	SmeI_131-246	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--R	PFKVVVFATDV	ITR-----	11545452
147	Atum_131-246	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--K	PFKVVVFATDV	ITPR-----	17741118
148	Core_163-278	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--R	PFKVVVFATDV	ITGR-----	13425051
149	Sep_509-626	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAQG--N	PFKVVVFANDV	ITTA-----	52010400
150	Psyr_385-500	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--L	PFKVVVFATDV	ITRQ-----	23468713
151	Psyr_385-500	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--L	PFKVVVFATDV	ITRQ-----	28852878
152	Bbac_288-403	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--K	PFKVVVFASDIT	ITQ-----	39576448
153	Naro_364-479	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--K	PFKVVVFASDIT	ITQ-----	48847895
154	Cvio_18-135	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--Q	PFKVVVFALDV	ITQE-----	34102765
155	CjeJ_24-137	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--Y	PFKVVVFANDIT	ITQR-----	6968544
156	Sep_18-132	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--T	PFKVVVFANDIT	ITKR-----	52011981
157	Mdeg_17-134	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--S	PFKVVVFANDIT	ITEV-----	48863742
158	Core_1-103	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--K	PFKVVVFANDIT	ITSE-----	13424441
159	CjeJ_142-259	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--K	PFKVVVFANDIT	ITSE-----	6968544
160	Agam_51-166	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--K	PFKVVVFANDIT	ITSE-----	31194935
161	Agam_116-231	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--K	PFKVVVFANDIT	ITSE-----	31195287
162	Vcho_145-260	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--V	PFKVVVFASDIT	ITDQ-----	9655902
163	Vcho_135-250	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--N	PFKVVVFASDIT	ITDK-----	28808720
164	Vcho_174-289	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--K	PFKVVVFASDIT	ITAE-----	9658296
165	Paer_139-254	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	53727981
166	Paer_139-254	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	9947925
167	Paer_20-135	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	53727683
168	Paer_3-112	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	9947371
169	Cvio_142-259	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	34103010
170	Ypes_136-253	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--N	PFKVVVFANDIT	ITEQ-----	21958470
171	Ypes_132-249	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--N	PFKVVVFANDIT	ITEQ-----	51590152
172	Psyr_20-135	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--K	PFKVVVFANDIT	ITEQ-----	28855718
173	Psyr_20-135	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--N	PFKVVVFANDIT	ITEQ-----	23471312
174	Pflu_18-135	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	48730703
175	Cvio_20-137	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	34103010
176	Cvio_140-257	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--V	PFKVVVFANDIT	ITEQ-----	34102765
177	Lmon_8-124	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	47013964
178	Lmon_8-124	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	46881199
179	Linn_8-124	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	16414310
180	Esp_14-130	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	46113970
181	Oihe_9-126	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	22776939
182	Esp_11-127	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	46113993

Generic PAS

183	Psyr_20-135	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	46188835
184	Psyr_20-135	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	15077778
185	Psyr_35-150	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	28853310
186	Pput_26-143	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	24982183
187	Pflu_18-135	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	48730078
188	Vpar_19-136	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	28808847
189	Psyr_22-139	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	46188225
190	Psyr_47-164	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	28851464
191	Psyr_17-135	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	28855155
192	Vcho_63-177	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	9654496
193	Sep_389-504	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	52010400
194	Psyr_89-209	LNK	-----G-----	EYDQGYKRI	KG	GVVWIIQATYNPIFDAHG--R	PFKVVVFANDIT	ITEQ-----	46188184

Figure C.1 continued

195	Psyr_88-209	LNDRSGN---T-----PY-R	IKN-RLAM-KN--GT	YRWFYAQGETLRDAR--GT	PLRVAGSLRDIHDE-----	28854917
196	Cvio_92-213	LN RGGT---T-----PY-D	IEY-RLQC-KN--GD	YRWFARFARGATLRDGK--GV	PLRVAGSLADITAI-----	34103227
197	Psyr_231-348	LNDRSGK---T-----PF-D	IEY-RLKM-KT--GE	YRWFARFARGQTRRNPE--GV	PLRVVGVGVVHLK-----	46188184
198	Psyr_227-348	LNDRSGK---T-----PF-D	IEY-RLKM-KT--GE	YRWFARFARGQTRRTPPE--GT	PLRVVGVGVVHLK-----	28854917
199	Cvio_231-352	LNDRSGQ---T-----PY-D	LDY-RLQC-KN--GE	YRWFARFARGQTRRAAD--GA	PLRVVGVGVVDAE-----	34103227
200	Rrub_41-162	LSOTTGR---T-----GY-D	VTY-RLKM-RD--GA	YRWFARFATGGCLRDAA--GK	PLRACGSLTVVHEQ-----	48766385
201	Rrub_173-294	LSOTTGR---T-----GY-D	VTY-RLKM-RD--GA	YRWFARFATGGCLRDAA--GN	PLRACGSLTVVDAV-----	48766385
202	Rpal_23-144	LSOTTGR---I-----CY-D	VKY-RLKV-KD--GS	YRWFARFATGGVLDEN--RK	PRRACGSLVDDEL-----	39648851
203	Rrub_28-149	LSRSAV---KD-----SY-D	VSY-RLKR-KD--GA	YHWFARFMGGVNRDGA--GK	ATRMCGSLVDIAE-----	48764228
204	Mmaz_29-147	IK---N-RT-----PY-Q	VEY-RIKK-AD--GS	TVFQEQAHFVNDDK--GN	LAYVDGVFLDVTOQ-----	20904696
205	Mace_31-149	IK---N-RT-----AY-Q	VEY-RIKK-TD--GS	TVFQEQLAHLVNDDA--GN	LAYVDGVFLDVTOQ-----	19917084
206	Mmaz_29-147	IK---T-RS-----PY-Q	VEY-RIQK-SC--GD	TVFQEQARLVNDEH--GN	IAYIDGVFLDVTOQ-----	20906162
207	Mbur_32-150	IK---N-NA-----Y-K	VNY-RIKT-KN--GD	TVYIREEGKLVNDEE--GN	AYLDGVFLNITEN-----	46142191
208	Mmaz_153-269	LE---T-GE-----GVFN	LR-ALKL-IAQ--DK	PLHTVTSVPIKDDT--GAI	VANLTIIDMTM-----	20904696
209	Mace_155-271	LE---T-GE-----AVYN	LER-PLKL-RAL--DK	PLHTVTSVPIKDDS--GAI	IGNLTIIDMTM-----	19917084
210	Mmaz_153-269	NE---H-E-----CIY	EK-SIKF-KAL--DK	PLFTVLSAVPVKDET--GT	AGSLMVIIDMTM-----	20906162
211	Mbur_156-284	LD---T-RE-----SVES	LEA-ITKL-HGL--D	ELYTISSASPIFDEE--GV	FEGILEVITDLDI-----	46142191
212	Aful_169-283	LD---T-GK-----AVIN	HQA-VTKT-KD--GR	EVPLVNSCIPVY--VD	GEMVGVLDLFDITEL-----	2649560
213	Aful_189-305	IK---T-RE-----RIEN	VEV-KLVV-KD--GS	FIASASIPVY--VG	DEFAGYIEVYDITEL-----	2649548
214	Aful_62-177	FE---N-KM-----LIEG	KEG-FLGV-KT--GK	AMPILTICAPVY--VD	GEFEGMVDFFDITEQ-----	2649548
215	Aful_35-157	LE---K-GG-----LIEH	QEV-RLGI-EK--G	LMHILTSCAPVK--VN	GLVGMGVFYVDVTP-----	2649560
216	Hsp_8-123	LS---R-GE-----AIRE	TSV-RTSE-LP--AC	AQHAKASATPLH--ND	GAVIGAVEVLITVDV-----	10580941
217	Aful_319-440	LD---N-PHEAHKLYDVIRKHPV	EGAYLTEI--NLNFPNRN	GRKAYVRATAAPVYNEK--GE	IGVVESTIEDITE-----	2649548
218	Ddes_384-499	IK---V-N-----GPVK	ADI-MSID-RD--GC	LYVKPSADVLRDAQ--GR	KMGYVEVASVDTI--V-----	53691249
219	Dvul_426-541	NT---K-Q-----GRYE	AET-IEIN-IH--GK	SRWIRPYGDILHDC--GQ	RAGYLEVASDVT-----	46448762
220	Dpsy_501-615	IK---T-D-----SIIT	DNT-IARP-QD--GI	IPIKYTGAPIKDAK--GN	KGLEYILDVTEE-----	50877249
221	Dpsy_379-493	IK---T-D-----SIIT	DNT-IARP-QD--GI	IPIKYTGAPIKDAK--GN	KGLEYILDVTEE-----	50877249
222	Dpsy_257-371	IK---T-D-----SVIT	DHT-IARP-QD--GI	IPIKYTGAPIKDAK--GN	KGLEYILDVTEE-----	50877249
223	Dpsy_135-249	IK---T-D-----SVIT	DHT-IARP-AD--G	IPIKYTGAPIKDAK--GN	KGLEYILDVTEE-----	50877249
224	Dpsy_623-736	IK---T-D-----SVVD	DRT-TASP-N--G	EMSIKYTGSPKDAK--GN	KGLEYILDVTEA-----	50877249
225	Dvul_393-508	LE---D-GK-----PHEN	VPE-SYTS-TV--GE	FEMIDVMPIRDAS--GD	IGGITFWNDVTEL-----	46448995
226	Dvul_261-373	IK---G-GN-----NQCG	N-H-SYRR-DD--G	VFLHYEVSPLRDDR--GAV	NGAIVLIDTQE-----	46451007
227	Ddes_360-475	MA---E-KQ-----VISN	IEV-IITG-HK--GF	TEVLNAVNTYLDME--GT	ITGMCILYDMEF-----	53691138
228	Dvul_359-474	KN---E-RK-----AITN	IDV-IHK--G	EQVQLNAVNTYLDTD--GN	IFGGFCLYDTEA-----	46449695
229	Dvul_118-232	LE---Q-GT-----IAR	REV-DLVG-RK--GK	RRLHINASPLYDLD--GT	LMGALCIYQDTEL-----	46448495
230	Ddes_118-232	LE---H-HT-----VTS	REV-DLIT-RK--GN	TRRIQIHASPLFDLS--GGL	MGALCIYQDTEL-----	53691151
231	Dvul_384-499	LR---E-GC-----AERN	IEV-DLT--SK--GE	TVHTLVDAVPLSDLD--GSL	IGSTFIYADITAI-----	46450563
232	Ddes_263-378	MA---N-RK-----AITG	VEV-EFT--RD--G	TRYSLDSSPIYDLD--GN	LMGAFITVCTDITE-----	23473878
233	Dvul_366-480	IK---S-RT-----DLHD	VA-VWNA-PS--GR	EVHLNVATTPFYDMD--GEL	LGSIAFVMDITDI-----	46450119
234	Dvul_389-504	LE---T-GE-----QIEA	LER-EMRD-TS--GH	VRRVLDAAPLNLDL--GSL	IGATIALIADITVI-----	46449710
235	Ddes_379-494	LE---S-GE-----TVTN	VER-EWRT-EK--GM	MHRVIDAAPLHDL--GR	VIGAILVADITDI-----	53691131
236	Dvul_405-518	IS---H-GK-----TQ	TEE-NLR--KK--G	LRTARLTAAPLHRRD--GN	IGAVLFLDITDI-----	46448526
237	Xcam_140-257	MR---A-L-----SKAVRA	PF-FG--RQ	IDFVYSPIIAAD--GT	KLGTIAQMMDVTAQ-----	21114304
238	Xcit_156-271	MR---A-LN-----SKAVRA	PF-FG--RQ	IDFVYSPIIAAD--GT	KLGTIAQMMDVTAQ-----	21109459
239	Xcam_163-285	LQ---Q-LT-----A	VSERY-TF--GP	VLDTMTPLYAGD--GR	SGSMLLRVNSAEVS-----	21112986
240	Xcit_164-284	LR---Q-LQ-----A	VSERY-TF--GP	VLDTMTPLYAPP--GR	SGSMLLRNISAE-----	21108110
241	Xcam_145-249	VE---S-GA-----G	AR-AL-GA--LQ	QESFVPHDHD--GT	PLGVVWVDRQSLILLEAQ--TT-----	21111298
242	Xcam_277-395	MA---A-LD-----R	DGTTTFEE-RF--G	AVFAQTVTTIQDED--GQ	VGDVCEWRDITE-----	21112973
243	Neur_277-373	ID---A-LD-----N	RI--GE--RT	YSLLLMPTV--ES--GER	AVVWVDRDTEQ-----	30180852
244	Vvul_42-160	LS---T-PA-----N	LPY-STVI-AI--DD	VRLEINVGAMLDAA--GN	VGNTEWQDTE-----	37201894
245	Vvul_38-156	LS---T-PA-----N	LPY-STVI-AI--DD	VRLEINVGAMLDAA--GN	VGNTEWQDTE-----	27359118
246	Vvul_38-157	LA---D-PS-----H	LPY-STVI-NI--KG	VKIELIVGAILDDR--GS	YGNTEWROVTEE-----	9658538
247	Vvul_169-288	LS---D-PN-----N	LPY-RTDI-AV--GD	IRIELNVAAVKNK--GE	YIGNSLEWROVTEQ-----	37201894
248	Vvul_165-284	LS---D-PN-----N	LPY-RTDI-AV--GD	IRIELNVAAVKNK--GE	YIGNSLEWROVTEQ-----	27359118
249	Ssp_146-263	LS---D-PK-----N	LPE-KTDI-TV--GE	MKFALNVDFIDDD--GT	VGNLEWADVTEA-----	52010400
250	Vvul_296-413	LS---N-PD-----R	LPE-SSDI-KV--GS	LEFNLTCIAMRDGS--GN	YMGFALQWIDITEQ-----	37201894
251	Vcho_164-282	LS---N-PE-----R	LPE-TSMI-KV--GS	LEFNLTCIAMRDTK--GE	YIGFALQWIDITEQ-----	9658538
252	Rrub_9-120	MA---K-PG-----A	LPH-HAVI-AL--GD	EFLDLQIEAL--GER	AAPKAVLTWSIVTER-----	48765139
253	Rrub_130-241	LS---K-K-----R	LPE-HSKI-RL--GP	ETLDRVTAIFGER--GE	YAMLCLWSVSTHL-----	48765139
254	Mmag_261-371	MQ---D-PT-----K	LPH-VANI-SL--GQ	EVIELNVSAILDRK--GH	YGPLLTWMPVTEK-----	46201816
255	Mdeg_138-254	LR---E-LK-----T	P-Y-KTTL-NI--GD	MVFGLIATPWFNSN--GER	LGTLVENLOKTEE-----	48863724
256	Rpal_13-128	IK---S-LA-----S	V-H-RATI-QI--G	RIFDLIATPINNAD--GS	RAGVVWVNDASIR-----	39648804
257	Sone_145-261	ID---K-LD-----R	K-Y-ESQI-QV--AS	CHFFLTASPIILTS--GER	LGSVVWVLDRTTE-----	24348039
258	Mdeg_272-640	LS---K-LK-----S	T-Y-STQI-KV--GI	RTESLIANPIKDD--GER	VGVWVWVLDRTTE-----	48863731
259	Mdeg_269-384	LN---N-LT-----S	T-Y-NGGA-KV--GG	RSETVIANPI--VD--GK	IGAVVWVLDRTAE-----	48863724
260	Cvio_457-573	LA---N-LR-----T	RAEL-QV--AG	RTFSLVANPVFDAD--GER	LGSVVWVLDRTAE-----	34332882
261	Paer_173-289	LA---N-LT-----G	V--KAEI-NL--GG	RRFSLDVVPVFNDAA--NER	LGSVQVLDRTTE-----	53726991
262	Paer_173-289	LA---N-LT-----G	V--KAEI-NL--GG	RRFSLDVVPVFNDAA--NER	LGSVQVLDRTTE-----	9946009
263	Daro_436-552	LE---R-LT-----G	T-H-RATI-RL--GG	RVFALTVPVINTR--GG	RLGFVWVLDRTNE-----	53729524
264	Sone_21-137	LE---R-LT-----Q	S-H-TAQI-SI--GK	RIFKLITLPIISRD--NK	HGTGVWVLDRTES-----	24348039
265	Daro_261-377	LG---T-LR-----G	I-H-RTEI-DI--GG	RYFSLVACPIVNDQ--GER	HGTGVWVLDRTAE-----	53729525
266	Cvio_181-297	MQ---Q-VR-----E	S-H-RSSI-SV--GG	RTFGLILTPILGAK--GER	LGDVWVLDQNTM-----	34332882
267	Cvio_319-435	LQ---Q-AR-----G	T-H-RSSI-VV--GG	RTFGLILSPIFNDR--GD	RLGAVVWVLDNTAE-----	34332882
268	Cvio_43-159	MQ---Q-LT-----G	T-H-RGTI-KV--GG	RTFSLVLTPIRDGQ--NR	KLGAVVWVLDITRD-----	34332882
269	Lint_219-335	LD---K-LS-----D	T-F-RSSI-TI--GG	REFDLIANPIVDVN--GN	KLGTVWVWSDVTEQ-----	24196190
270	Lint_219-335	LD---K-LS-----D	T-F-RSSI-TI--GG	REFDLIANPIVDVN--GN	KLGTVWVWSDVTEQ-----	45600635
271	Lint_349-466	LD---L-LT-----D	T-F-RSSI-NI--GG	RTFNLIANPIIDET--GD	RLGSVVWSDVTEQ-----	24196190
272	Lint_481-597	LS---S-PT-----G	I-H-KATI-KI--GG	RTFDLIANPIILDSN--GK	RLGSVVWSDVTNE-----	24196190
273	Cvio_19-144	LA---DLASG-----K	I-PQHTAMI-FV--GE	VMFETHVFPVWDSANPSQL	LCFMASFDVSSE-----	34103223
274	Gsul_12-134	LG---KPG-----E	MP-HSAEI-PI--GG	ITLRTSFPPVWDSKNPGR	VKCYMACVDDITAE-----	39985192

Figure C.1 continued

PAS_Aer

1	Daro_8-120	--TDVETRLPEGOFLYSRTD--LKGITLNEAEFAQIS-----AY--RREELIGEN--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
2	Daro_4-120	PVTNVEHLPEGEFLYSSTD--LQGNLVEANAFAKIS-----NF--SREELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
3	Reut_6-119	--TD EYRPSDEVLITRTD--AQGNLVEANAFRRSS-----GY--DRAELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
4	Rgel_11-124	--AACTVSAYQOAPLITRTD--LQGNLVEANAFRRSS-----GY--AMEQLVIGAP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
5	Bcep_8-121	--TQREFFDFPDATLMSTTD--ANSYITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
6	Bcep_8-121	--TQREFFDFPDATLMSTTD--ANSYITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
7	Bfun_8-121	--TQREFFDFPDATLMSTTD--ANSYITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
8	Rsol_8-121	--TQREFFDFPDATLMSTTD--ANSYITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
9	Bcep_8-121	--TQREFFDFPDATLMSTTD--ANSYITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
10	Bcep_8-121	--TQREFFDFPDATLMSTTD--ANSYITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
11	Bmal_1-101	-----MSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
12	Bmal_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
13	Bmal_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
14	Ecol_8-121	--TQNTPLADDTLMSTTD--LQSYITHANFVQVS-----GY--TLQELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
15	Ecol_8-121	--TQNTPLADDTLMSTTD--LQSYITHANFVQVS-----GY--TLQELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
16	Styp_8-121	--SOLNTEPLDDTLMSTTD--LESYITHANFVQVS-----GY--OLNELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
17	Styp_8-121	--SOLNTEPLDDTLMSTTD--LESYITHANFVQVS-----GY--OLNELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
18	Ecar_8-120	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
19	Ypes_8-120	--SRQYPIERDITLQSTTD--IHGNIYANAFVQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
20	Ecar_8-121	--SOLNTEPLDDTLMSTTD--LESYITHANFVQVS-----GY--OLNELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
21	Ecar_8-118	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
22	Reut_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
23	Rmet_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
24	Avin_10-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
25	Rsol_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
26	Bbro_8-121	--YDRTKREDQVLSRTD--TKGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
27	Tden_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
28	Bjap_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
29	Rpal_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
30	Daro_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
31	Neur_4-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
32	Mfia_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
33	Tden_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
34	Rgel_8-120	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
35	Bfun_4-120	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
36	Bbro_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
37	Lint_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
38	Psyr_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
39	Psyr_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
40	Pput_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
41	Pfli_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
42	Paer_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
43	Psyr_8-122	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
44	Psyr_8-125	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
45	Pput_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
46	Pput_8-120	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
47	Pput_8-120	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
48	Vvul_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
49	Vcho_8-120	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
50	Vpar_4-117	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
51	Ppro_1-115	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
52	Vvul_5-119	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
53	Sone_6-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
54	Sthe_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
55	Sone_14-127	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
56	Ecol_6-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
57	Vcho_11-124	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
58	Pfli_6-123	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
59	Vpar_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
60	Vvul_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
61	Vcho_24-136	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
62	Ppro_6-119	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
63	Ppro_8-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
64	Sone_6-119	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
65	Rpal_7-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
66	Rpal_6-120	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
67	Rpal_7-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
68	Mmag_1-109	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
69	Bjap_7-122	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
70	Rpal_1-124	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
71	Rpal_20-133	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
72	Mmag_8-122	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
73	Wsuc_9-123	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
74	Wsuc_7-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
75	Wsuc_7-121	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
76	Daro_1-109	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
77	Cje_1-113	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----
78	Cvto_9-125	--TQREFFDFPDATLMSTTD--PHGRITVANAFAIQVS-----GF--SPEELIGQP--HNMVR-----HPDMP--AAAFADNMWDLRAGR-----

PAS_Che

79	Ypes_15-131	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
80	Ypes_1-112	-----MISLNAVPMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
81	Ypes_11-127	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
82	Paer_17-134	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
83	Paer_17-134	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
84	Vpar_13-130	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
85	Vcho_53-169	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
86	Vcho_23-140	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
87	Pput_23-138	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
88	Pfli_23-138	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
89	Agam_1-111	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
90	Pput_148-265	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
91	Pfli_140-257	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----
92	Psyr_145-261	--PRAELTSIDNAVEMILFK--PDGTIVQVNLFLAAM-----GY--QKDEVLGKH--HKIFC-----DPQYA--SDAYRRHWQLNENGR-----

Figure C.2 Alignment of chemotaxis PAS domains after manual editing and visualized with VISSA.

```

93 Psyr_170-286 --ENSAFIQALLRSTAVIEFD--LSGHVLTANQFLRGM-----GY--NLAQIKGKH--HSLFC-----DPAETSLAPYREFWALNRGE-----
94 Psyr_140-257 RENEALVNALQRSTAVIDFT--LDGTVISANNEFLRAM-----GY--ELNQIKGKH--HKMFC-----VPEESASDAYTQFWERLRGE-----
95 Psyr_155-272 RENEALVNALQRSTAVIEFT--LDGMVITANDNEFLKAM-----GY--ELNQIKGKH--HKMFC-----VPEESNADAYSHEFWERLRGE-----
96 Vpar_141-258 REYEDMLNALS-SHAVIEFT--LDGTVLKANDNEFLSTM-----GY--KHEQIVGKH--RIFC-----LPEEANSAYHDFWKRILASGQ-----
97 Psyr_125-242 QESIELLCLNRS-IAVQES--LDGVLVANPEFETM-----GY--SLAEIKGKH--HRLFC-----LDEDAASAEYAKFWKSLSGV-----
98 Psyr_140-257 AENEALIDALRSTAVIQEN--LDGTVITANQOFLQAM-----GY--TLQAVNRS--RMFC-----HAGDEASPOYTAFWKKLNQGE-----
99 Pput_143-260 HESESLKAIKRSMAVIEFT--PQGVVIRKANNEFLTOM-----GY--RLDEVVGRH--GLFC-----LAHERESAOYREFWASLNQGE-----
100 Pflu_143-260 KEESMLAAIGRSMVIEFT--PEGNVITANDNEFLTKM-----GY--SLNEIVGHH--HSLFC-----HRVEASSAQYKAFWASLNQGE-----
101 Psyr_144-257 ---SKLAAVIRAN-ACEFE--PNGNVITANNEFLNVM-----GY--ALAEIKGKH--HSLFC-----EPTLVNSPEYTEFWKKLNQGE-----
102 Psyr_143-257 ---NSKLAAVIRAN-ACEFE--PNGNVITANNEFLNVM-----GY--ALAEIKGKH--HSLFC-----EPTLVNSPEYTEFWKKLNQGE-----
103 Paer_140-257 HEMQSKLDALS-SH-AMIEFD--LDGMVITANDNEFLATM-----GY--GRAELASAN--HRQFC-----EPGYRDCPOYADLWRRLNQGE-----
104 Paer_117-234 HEMQSKLDALS-SH-AMIEFD--LDGMVITANDNEFLATM-----GY--GRAELASAN--HRQFC-----EPGYRDCPOYADLWRRLNQGE-----
105 Pflu_144-257 ---AKLQALDRAN-AVIEFD--LDGMVITANDNEFLTRM-----GY--TLAEIKGKH--RLFC-----PAQLVNSAYQDFWRRLNQGE-----
106 Bjap_134-256 REASAKVSAIS-AQAVIEFK--LDGTVITANNEFLKAL-----GY--SLAEIKGKH--HSLV-----AQSERDCGAYREFWAAKNQGE-----
107 Rpal_135-250 --DAGKLAATIGRAQAVIEFA--MDGTILTANNEFLAAM-----EY--SLGELNGRH--SMFV-----EPSVRESDDYREFWALNRGE-----
108 Bjap_263-378 --LAGQIAAIDKAQAVIEFN--MDGTILTANNEFLGTI-----GY--SLAEIKGHH--HSMFV-----EPAERDCGAYREFWAAKNQGE-----
109 Rpal_256-371 --MAGQIAAIGKSAQVIEFD--MDGTILTANNEFLRAL-----GY--SLAEIKGQP--HSMFV-----DPSERAGAYREFWASLNQGE-----
110 Rpal_255-372 A-FSGQIDAIRKSAQVIEFS--IDGTVLIDANNEFLHAL-----GY--SLGELNGRH--HSMFI-----DPAERDCGAYREFWAAKNQGE-----
111 Rpal_133-249 --DFAGQVAAA-RSQAQVIEFN--MDGTIRTANNEFLKTI-----GY--TLDEIKGEH--HNMFV-----EPADRDCGAYREFWAAKNQGE-----
112 Bsp_90-205 --LLKGVNAIEKSAQVIEFK--LDGTILTANNEFLNVL-----GY--SLGELNGRH--HSMFV-----DPAERDCGAYREFWAAKNQGE-----
113 Rpal_11-128 DHAAAMLAALNRSQAQVIEFD--LDGMVIDANDNEFLTAL-----GY--SLPEIKGKH--HRMFV-----DPSHHSTAYREFWALIRAGQ-----
114 Bjap_19-134 --ADAQALAAIGRSMVIEFA--MDGTILTANNEFLKAI-----GY--SLDEIKGKH--HAMV-----PADQRDCGAYREFWAAKNQGE-----
115 Rub_90-205 --LOEKVAAISRSQAQVIEFA--MDGTILTANNEFLTKM-----GY--RLDEIQGN--HSLFI-----APGERDCGAYREFWAAKNQGE-----
116 Mmag_103-220 AFEYEGQIAAINRSQAQVIEFK--PDGTILTANNEFLNAI-----GY--RLDEVVGRH--HSLV-----DAEEVRSFAYAEFWALIRAGQ-----
117 Xcam_138-255 A-FEGRIDAIDKVA-IVIEFS--LDGTVLIDANNEFLVVM-----GY--RLDEIRGQH--HRLFV-----EPAGRDCGAYREFWAAKNQGE-----
118 Xcit_161-278 A-FEGRIDAIDKVA-IVIEFS--LDGTVLIDANNEFLVVM-----GY--RLDEVVGRH--HSLV-----DAGTROSGAYREFWAAKNQGE-----
119 Xcam_262-377 --ADGRITAI-VNGVIEFD--LDGTVILANNEFLAIM-----GY--RADEAIGKH--SVFV-----DAAYASSEDYREFWALIRAGQ-----
120 Xcit_285-400 --ADGRITAI-VNGVIEFD--LDGTVILANNEFLAIM-----GY--PAEAIGQH--HSLFV-----DAAYASSEDYREFWALIRAGQ-----
121 Xcam_18-133 --LOHKATAM-RVMVIEFD--LEGRVILANDNEFLAM-----GY--RLDEVVGRH--HRMFV-----HPSERESDGYREFWALIRAGQ-----
122 Xcit_41-156 --LRHKVAAVDRVMVIEFD--LDGRVILANDNEFLAM-----GY--RLDEVVGRH--HRMFV-----TAADRDCGAYREFWAAKNQGE-----
123 Atum_8-124 ANACAVLAALSKSQAMIEFD--LTGRILTANNEFLCAL-----GY--ELAEIKGKH--HSMFV-----EPDVRSSADYKAFWAKLAGN-----
124 Atum_13-129 ANACAVLAALSKSQAMIEFD--LSGRILTANNEFLCAL-----GY--ELSEIKGKH--HSMFV-----EPAPVRSADYKAFWAKLAGN-----
125 Atum_10-124 --AVAVLAAL-KSQAMIEFD--LSGKILTANNEFLCAL-----GY--ELAEIVGRH--HSLFV-----EPSVSSPDYKAFWAKLAGN-----
126 Atum_9-125 LIDASAVLDAIRKSAQVIEFD--LTGKILKANDNEFLKAV-----GY--QPEIVGRH--HSLFI-----SSEDAASPEYKAFWAKLAGN-----
127 Sent_8-124 RDAQOIDAIRKSAQVIEFD--LAGNVILANDNEFLCAL-----GY--SLQEVIGQH--HSLFC-----APEFVATEYREFWAAKNQGE-----
128 Atum_8-124 ADNNMMLDAIRKSAQVIEFD--LKGILTANNEFLCAL-----GY--DLAEIKGKH--HSLFI-----DRETAASHAYQEFWESLAGN-----
129 Ccre_43-156 ---QKIDALDKS-AMIEFD--VKGILANNEFLCAL-----GY--EAREIKGKH--HSLFV-----DPEYASQAYREFWALIRAGQ-----
130 Sep_21-138 GEQOAMLDVSRVH-AMIEFE--LDGTIRTANNEFLKVI-----GY--QLEIKGKH--HRMFC-----DDPVVASTYKDEFWALIRAGQ-----
131 Psyr_141-256 --YEGKVAIDRSQOIEFD--LNGRVILANDNEFLKVI-----GY--RLDEIQGN--HSLFV-----DDPVVASTYKDEFWALIRAGQ-----
132 Psyr_141-256 --YEGKVAIDRSQOIEFD--LNGRVILANDNEFLKVI-----GY--RLDEIQGN--HSLFV-----DDPVVASTYKDEFWALIRAGQ-----
133 Psyr_263-378 --SAGKVTAIERSQAQVIEFD--LTGKVLHANNEFLAVE-----GY--DLDEVVGRH--HRMFC-----SEEFVSSLOYREFWALIRAGQ-----
134 Psyr_263-378 --YAGKVTAIERSQAQVIEFD--LTGKVLHANNEFLAVE-----GY--DLDEVVGRH--HRMFC-----SEEFVSSLOYREFWALIRAGQ-----
135 Psyr_18-134 --DHSILMTAIDRSQAMIEFD--LDGMVILANDNEFLCAL-----GY--RLDEVVGRH--HRMFC-----TPEHASSEVYREFWAAKNQGE-----
136 Psyr_21-134 ---SLMVAI-RSQAMIEFD--LEGNILNANNEFLCAL-----GY--RLDEVVGRH--HRMFC-----TPEYASSEVYREFWALIRAGQ-----
137 Naro_242-357 --CSEQIAAIS-SHAMIEFD--LEGNILNANNEFLCAL-----GY--RLDEVVGRH--HRMFC-----TPEYASSEVYREFWALIRAGQ-----
138 Vcho_184-299 --RSQMNAVNLTQAVIEFT--LDGTILTANNEFLQTV-----GY--QLEIKGKH--HSLFV-----DEQYKQSQEQYHFWALIRAGQ-----
139 Ccre_20-135 --LEGIVAATIRKSAQVIEFN--LDGSIITANNEFLRAV-----GY--GLSEIQGN--HSLFV-----DPAFPAAGAYREFWALIRAGQ-----
140 Bbac_38-159 NSIREETIKALHRVQAQVIEFN--LDGTILTANNEFLKTS-----GY--SLDEIQGN--SMFC-----EPEYGMSTYKQFWALIRAGQ-----
141 Bbac_165-281 --YEGKVTAIERSQAQVIEFN--LDGTILTANNEFLATV-----GY--GLSEIQGN--HSLFV-----DPAFPAAGAYREFWALIRAGQ-----
142 Sep_264-381 RLVAGVNLALDRSQAMIEFK--LDGTILTANNEFLATV-----GY--GLSEIQGN--HSLFV-----DPAFPAAGAYREFWALIRAGQ-----
143 Atum_131-246 --DAGKTEALSRQAQVIEFT--PTGDLTANNEFLSAL-----GY--SLSEIQGN--SMFC-----EPSYATSDYKQFWALIRAGQ-----
144 Atum_131-246 --DAGKTEALSRQAQVIEFT--PTGDLTANNEFLSAL-----GY--SLSEIQGN--SMFC-----EPSYATSDYKQFWALIRAGQ-----
145 Atum_132-247 --DDGKLAALSRQAQVIEFT--PDGKILKANDNEFLCAL-----DY--TAEIKGKH--HSLFC-----EPAYASQEDYREFWALIRAGQ-----
146 Smel_131-246 --DAGKTEALSRQAQVIEFT--PDGKILKANDNEFLCAL-----DY--TAEIKGKH--HSLFC-----EPAYASQEDYREFWALIRAGQ-----
147 Atum_131-246 --DAGKTEALSRQAQVIEFT--PDGKILKANDNEFLCAL-----DY--TAEIKGKH--HSLFC-----EPAYASQEDYREFWALIRAGQ-----
148 Ccre_163-278 --RTAKLDAVERVQAVIEFT--VDGVLNANNEFLATV-----GY--ALSEIQGN--HSLFV-----DPAEARSAYAEFWALIRAGQ-----
149 Sep_509-626 QDFNCKLEAISNASAVIEFT--PDGKILKANDNEFLCAL-----GY--SLSEIQGN--HSLFV-----DPAEARSAYAEFWALIRAGQ-----
150 Psyr_385-500 --DQGVNAIDRSQAQVIEFD--MAGNILTANNEFLKAI-----GY--GLSEIQGN--HSLFV-----DPAEARSAYAEFWALIRAGQ-----
151 Psyr_385-500 --DQGVNAIDRSQAQVIEFD--MAGNILTANNEFLKAI-----GY--GLSEIQGN--HSLFV-----DPAEARSAYAEFWALIRAGQ-----
152 Bbac_288-403 --DAGKISAIKGAQVIEFN--MDGTILTANNEFLKTV-----GY--GLSEIQGN--HSLFV-----DPAEARSAYAEFWALIRAGQ-----
153 Naro_364-479 --YQAKVVAIRNRAVIEFD--LDGMVITANDNEFLSTM-----GY--SLSEIQGN--HSLFV-----DPAEARSAYAEFWALIRAGQ-----
154 Cvio_18-135 ENKQNVLDALNOSTAIEFS--PDGKILSANNEEOTL-----GY--SAEELIGKH--HSLFV-----LPQVTASQEKYAEFWALIRAGQ-----
155 Cjej_24-137 ---DILRSIGNTMAVIEFT--TDGVILANNEFLTAM-----KY--SLSEIKGKH--HSLFV-----LPEVNSAYSDYREFWALIRAGQ-----
156 Sep_18-132 ---QKIRAMEN-NTQAVIEFK--VDGVLNANNEFLDIL-----GY--TLSEIQGN--HSLFV-----YPSVRSKDAYKQFWALIRAGQ-----
157 Mdeg_17-134 KELEKVNAAIRSAVIEFT--PEGVETANDNEFLNAN-----GY--SLNEIQGN--HSLFV-----TDEYRNSNDYKQFWALIRAGQ-----
158 Ccre_1-103 ---MLELR--FNAAVIRAN-ACEFILT-----GY--ASPEVIGCP--HSLFV-----AGESTDSQEKYAEFWALIRAGQ-----
159 Cjej_142-259 LDLRNTIAANRSMAVIEFK--PDGTILTANNEFLRAM-----DF--NIDEIKGKH--SMFC-----DSNYRHSKDYQFWALIRAGQ-----
160 Agam_51-166 MTQKAIIEALDRSIAVIEFE--PSGVITANNEFLTOM-----GY--RLDEIRGRQ--RMFC-----DELFP--YREHDFWQELAGT-----
161 Agam_116-231 LSQAITKALERSIAVIEFT--PNGDILSANNEFLSCL-----GY--TLAEIKGKH--HSLFV-----DDAF--YREHDFWQELAGT-----
162 Vcho_145-260 DQKAAVHSLDRSSAMIQFN--PDGTILKANNEFLKAT-----GY--RLSEIVGRH--HRMFC-----DRFP--YKENPFWALIRAGQ-----
163 Vpar_135-250 EQKEATLNALDLSLAVIEFD--REGILTANNEFLKTI-----GY--ELSDVQGN--HSLFC-----FDFP--YQENPFWALIRAGQ-----
164 Vcho_174-289 ESQRDLTIAL-QNFVIEFE--PDGTILSANNEFLTKM-----GY--SLDQIKGKH--RLFC-----FDEF--YQENPFWALIRAGQ-----
165 Paer_139-254 LILNAINDAIRKSAQVIEFT--PDGKILKANDNEFLRLF-----GY--SLKSIKQGN--RMLC-----FDEF--YQENPFWALIRAGQ-----
166 Paer_139-254 LILNAINDAIRKSAQVIEFT--PDGKILKANDNEFLRLF-----GY--SLKSIKQGN--RMLC-----FDEF--YQENPFWALIRAGQ-----
167 Paer_20-135 ---ERLHMAALDRSMARVIEFD--PDGMILTANNEFLTIL-----GY--RRDEILGKP--HRQLC-----DGAYASSEDYREFWALIRAGQ-----
168 Paer_3-112 ---ALDRSMARVIEFD--PDGMILTANNEFLTIL-----GY--RRDEILGKP--HRQLC-----DGAYASSEDYREFWALIRAGQ-----
169 Cvio_142-259 EDSRALROATDHSMAVIEFD--LDGMILTANNEFLKVF-----GY--RRADVIGKH--RMLC-----EPSYAGGPEYQOLWQELIRAGQ-----
170 Ypes_136-253 QEHQSILLEANRSMGMITFT--PQGITILANDNEFLNVI-----GY--SLADIQHS--HQLC-----LPEFAHSEYHQQWALIRAGQ-----
171 Ypes_132-249 QEHQSILLEANRSMGMITFT--PQGITILANDNEFLNVI-----GY--SLADIQHS--HQLC-----LPEFAHSEYHQQWALIRAGQ-----
172 Psyr_20-135 --LKGLTSALEKSMVAVELG--LDGKILRANNEFLATM-----GY--RADELTKNT--HRDFC-----EPEVLRSEYADLWASIKAGK-----
173 Psyr_20-135 --LKGLTSALEKSMVAVELG--LDGKILRANNEFLATM-----GY--RADELTKNT--HRDFC-----EPEVLRSEYADLWASIKAGK-----
174 Pflu_18-135 NQACGLLEANRSMVAVIEFD--VDGVLNANNEFLKTM-----GY--TREOVVGP--HRMFC-----SPEFARGNOYELWASIKAGK-----
175 Cvio_20-137 QORSLLAAT-RSMVIEFS--VDGKIVHANNEFLRAM-----GY--SADEILGKH--HSLFV-----QDYVANSAPYREFWALIRAGQ-----
176 Cvio_140-257 HEARNIMKAVERSMVIEFT--PEGILTANNEFLGAT-----GY--TAEELIGKH--HSLFV-----PADEVHSPAYEHLWQELIRAGQ-----
177 Lmon_8-124 EDATLLDGLQLQNAVIEFD--TNKKVITYANALFAEAM-----GY--SEEMIKLS--HPDLC-----FPDEVCSASYKAMVNLIRAGQ-----
178 Lmon_8-124 EDATLLDGLQLQNAVIEFD--TNKKVITYANALFAEAM-----GY--SEEMIKLS--HPDLC-----FPDEVCSASYKAMVNLIRAGQ-----
179 Linn_8-124 EDATLLDGLQLQNAVIEFD--TNKKVITYANALFAEAM-----GY--SEEMIKLS--HPDLC-----FPDEVCSASYKAMVNLIRAGQ-----
180 Esp_14-130 --DTQVIRALEONLAVIEFD--DRRVAVYNEFLATM-----GY--EYDELIGRY--HRD-----FPGFASPAYELWAKLAGN-----
181 Oihe_9-126 KMNLVVQALNESLAVIEFD--LTKVAVVNDNEAAL-----GY--KREELGMH--HSLFV-----FDTFVCSASYKAMVNLIRAGQ-----
182 Esp_11-127 LASADILEALTHN-AMIEFD--TQRRVVDVNDLAKTM-----KY--KREELGMH--HSLFV-----RADFANSASYKAMVNLIRAGQ-----

```

Generic PAS

```

183 Psyr_20-135 --LEQAGQILN-ETVAVVLD--GTGFIQTVN-LEFETEM-----SY--AQAEIVGRS--LSE-S-----PPELSGDVHQKRALTAIRGK-----
184 Psyr_20-135 --LEQAGQILN-ETVAVVLD--GTGFIQTVN-LEFETEM-----SY--AQAEIVGRS--LSE-S-----PPELSGDVHQKRALTAIRGK-----
185 Psyr_35-150 --LEQAGQILN-ETVAVVLD--RQKGQVTVNGLFETEM-----SY--SOSEIAGRS--LSEMS-----PPELSGDVHQKRALTAIRGK-----
186 Pput_26-143 M-LEQVKSLSDESEMLVLQLD--PQGIEMVNV-NFESEEM-----LY--RAEOLIGRN--TEDIV-----PAHVKS-LDFYKEMSAISIRGE-----
187 Pflu_18-135 SSILQVKSLESEMLVLQLD--PQGIEMVNV-NFESEEM-----LY--KSHDLIGRA--TEDV-----PAHVKS-LDFYKEMSAISIRGE-----
188 Vpar_19-136 YSILQVKSLESEMLVLQLD--PQGIEMVNV-NFESEEM-----LY--KSHDLIGRA--TEDV-----PAHVKS-LDFYKEMSAISIRGE-----
189 Psyr_22-139 SMYBOMKQMDAQVCLTLD--ASYHIVHANDNLLSTL-----GY--SLEQVIGKD--LDH-V-----PTYVKQIDCYRSKLVAVQKGE-----

```

Figure C.2 continued

190 Psyr_47-164 SMYRQMQRGM ARMVSLSLD--ASNRIAHAN:NFLRAL-----GY--TABQILGRE--LDOIV-----PTYVKQLDCYPNNLKLAIVQKE-----
191 Psyr_17-135 HVLHQLLGRI:QDMLTVQVD--GNFTISAN:QGFAKAL-----GY--TPDRISGFP--LSSIA-----AFDSKMPWFHGLKTTLLIFE-----
192 Vcho_63-177 QVNAQFVQVIR:HLALLECE--PNGTICYAS:AFAHLC-----RV--SAAEMVGAD--FANLW-----RTHQCPSVQRLLDQAKRGH-----
193 Sep_389-504 KTLISLVANETD--NSVLIAD--ADGREIYVN:GFTKLT-----GH--EYKDVIGKK--PGE L-----QGRHTT:FEFTRKRIRRENINAKQ-----
194 Psyr_89-209 --RASSEGLWDMDEVVAGDPV--NPNNRFWWSQOETRL-----GF--NDRER:PNVLASWADR:H-----PQ--DKOASLDAPAFKHINDRS--GNT
195 Psyr_88-209 --NRASSEGLWDMDEVVAGDPV--NPNNRFWWSQOETRL-----GF--NDRER:PNVLASWADR:H-----PQ--DKOATLDAPAFKHINDRS--GNT
196 Cvio_92-213 --NRASSEGLWDMDEVVAGDPV--NPNNRFWWSQOETRL-----GF--HHRER:PDVLGSWAGR:H-----PE--DRORVLDAFAAHLN:RG--GTT
197 Psyr_231-348 -----SDGLWDMDEVVAGDPV--NARNPFWWSQOETRL-----GF--EVEE:PDVLDSWASR:H-----PE--DKERSLTAFGAHLNDRS--GKT
198 Psyr_227-348 --RMLSDGLWDMDEVVAGDPV--NARNPFWWSQOETRL-----GF--EVEE:PDVLDSWASR:H-----PE--KERSLTAFGAHLNDRS--GKT
199 Cvio_231-352 --SEMLAEGLWDMDEVVAGDPV--SGAHAFWWSQOETRL-----GF--DEE:PDVLESMSR:H-----PE--DKOTVLDAFAAHLNDRS--GQT
200 Rrub_41-162 --DGNAGVGLWDAVIEGDPV--HAOSRWTWSAEETRLV-----GF--KSETE:PNVGSWADR:H-----PE--DRASTFEAFGALLSOTT--GRT
201 Rrub_173-294 --SHHAGVGLWDAVILHNG:AM--HPOSRTWWSPEETRL-----GF--ESEAE:PDVVRMSDR:H-----PD--VGPTFAAGGALLT:RS--GRT
202 Rpal_23-144 L:LHCGIGLWDAIIEGDPV--HFKARWWSPEETRL-----GY--SSEAE:PNVQMSDR:H-----PD--DVATITFAAFTQT:GTGI--
203 Rrub_28-149 --TRAGVGLWDAVIEGDPV--HPLSVHWSPEETRL-----GF--TDAAE:PDVKSASER:H-----D--VATLGAFTWWSGRS--AVKD
204 Mmaz_29-147 KETSLLID:LPVTVFRISNE--SSWAIHHIGRSVEQLT-----GY--SKMDF:TR--TWSDLIC-----PE--DIPALNKVQKAKNRT--
205 Mace_31-149 KEVSSLLIDNLPITVFRISD--SSWAVRYSIKRNVQVLT-----GH--TKMDF:SQK--LWSVDIVY-----PE--DAPLIEEIVQKAKNRT--
206 Mmaz_29-147 IT:IGWLLIDNLPVTVFRISNE--SSWPICHYISKVEVLT-----GY--PAAEF:SRRLWSVDIVF-----PE--DVSRIIDANISMKNRT--
207 Mbur_32-150 EETSLLID:LPVTVFRISNE--ASWDLYYISKKNVYDIT-----GH--PKDFE:GKK--LWSVDIVF-----PE--DVPKIDAAIDAKNNA--
208 Mmaz_153-269 ESQKATVYSIPKPSLALYVD--ASGKIKYINDYFVKM-----KF--KSAEAGLS--PADLM-----ESNN:KSAIAETVLEIG--GV
209 Mace_155-271 ESQKATVYSIPKPSLALYVD--ASGKIKYINDYFVKM-----KF--KSAEAGLS--PADLM-----ESNN:KSAIAETVLEIG--GV
210 Mmaz_153-269 ESQKATVYSIPKPSLALYVD--ASGKIKYINEYFLTIC-----KF--KSAEAGLS--PADLM-----ESNN:KSAIAETVLEIG--GV
211 Mbur_156-284 ESQKATVYSIPKPSLALYVD--ASGKIKYINEYFLTIC-----KF--KSAEAGLS--PADLM-----ESNN:KSAIAETVLEIG--GV
212 Aful_169-283 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
213 Aful_189-305 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
214 Aful_62-177 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
215 Aful_35-157 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
216 Hsp_8-123 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
217 Aful_319-440 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
218 Ddes_384-499 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
219 Dful_426-541 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
220 Dpsy_501-615 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
221 Dpsy_379-493 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
222 Dpsy_257-371 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
223 Dpsy_135-249 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
224 Dpsy_623-736 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
225 Dvul_393-508 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
226 Dvul_261-373 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
227 Ddes_360-475 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
228 Dvul_359-474 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
229 Dvul_118-232 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
230 Ddes_118-232 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
231 Dvul_169-288 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
232 Ddes_263-378 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
233 Dvul_366-480 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
234 Dvul_389-504 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
235 Ddes_379-494 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
236 Dvul_405-518 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
237 Xcam_140-257 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
238 Xcit_156-271 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
239 Xcam_163-285 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
240 Xcit_164-284 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
241 Xcam_145-249 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
242 Xcam_277-395 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
243 Neur_277-373 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
244 Vvul_42-160 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
245 Vvul_38-156 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
246 Vcho_38-157 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
247 Vvul_169-288 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
248 Vvul_165-284 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
249 Ssp_146-263 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
250 Vvul_296-413 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
251 Vcho_164-282 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
252 Rrub_9-120 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
253 Rrub_130-241 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
254 Mmaz_261-371 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
255 Mdeg_138-254 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
256 Rpal_13-128 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
257 Sone_145-261 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
258 Mdeg_527-640 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
259 Mdeg_269-384 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
260 Cvio_457-573 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
261 Paer_173-289 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
262 Paer_173-289 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
263 Daro_436-552 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
264 Sone_21-137 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
265 Daro_261-377 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
266 Cvio_181-297 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
267 Cvio_319-435 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
268 Cvio_43-159 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
269 Lint_219-335 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
270 Lint_219-335 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
271 Lint_349-466 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
272 Lint_481-597 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
273 Cvio_19-144 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A
274 Gsul_12-134 ELVKEVFRNMPFPAYVLEVN--RDHKIYQANDEAKLA-----GY--SSAETI:GLP-----GTSGGT:TVADKVID GK--A

PAS Aer

1 Daro_8-120	PWRGVVKNRRRDG:YYVVLANASPI:EH-G--QIVGYQSVRT	PGR-	46140309
2 Daro_4-120	PWRGVVKNRRKDDG:FYVVLANASPV:EH-G--QVVGYSQSVR:RPSR-		53729639
3 Reut_6-119	PWTGVVKNRRKDDG:FYVVLANITPVQD-G--KPSGYLSVRT:APTKA		53762384
4 Rgel_11-124	PWTGLVKNRRSSSG:AFVVKANI:IPMKFD-R--QTVGFTSVQC:PPADA		47573623
5 Bcep_8-121	PWTALVKNRRKNGD:HYVVRNATPVVVRN-G--OPTGYMSVRT:KASRD		46322895
6 Bcep_8-121	PWTALVKNRRKNGD:HYVVRNATPVVVRN-G--EPKGYMSVRT:KATRD		46317934
7 Bfun_8-121	PWSALVKNRRKNGD:HYVVRNATPVVVRN-G--OPAGYMSVRT:QASRE		48783787
8 Rsol_8-121	PWSALVKNRRKNGD:HYVVRNATPVVVRN-G--RPAGYMSVRT:KPTRD		17431698
9 Bcep_8-121	PWTALVKNRRKNGD:HYVVRNATPVVVRN-G--EPQGYMSVRT:KAPHD		46317933
10 Bcep_6-121	PWTALVKNRRKNGD:HYVVRNATPVVVRN-G--APHGYMSVRT:KAPRD		46322894

Figure C.2 continued

11 Bmal_1-101	PWTALVKNRRKNGDHYWVRANAVPVIRG-G--QTQGYMSVRT PARA	52423246
12 Bpse_8-121	PWTALVKNRRKNGDHYWVRANAVPVIRG-G--QTQGYMSVRT PARA	52211725
13 Bmal_8-121	SWTAVIKNRRKNGDHYWVRANATPVIRN-G--QLVGYMSVRTKPSRE	52428136
14 Ecol_8-121	PWSGIVKNRRKNGDHYWVRANAVPMVRE-G--KISGYMSIRTRATDE	26110084
15 Ecol_8-121	PWSGIVKNRRKNGDHYWVRANAVPMVRE-G--KISGYMSIRTRATDE	13363427
16 Styp_8-121	PWSGIVKNRRKNGDHYWVRANAVPMIRE-G--RVTGYMSIRTRATDD	16421774
17 Styp_8-121	PWSGIVKNRRKNGDHYWVRANAVPMIRE-G--RVTGYMSIRTRATDD	16504293
18 Ecar_8-120	SWTGLVKNRRNNGDHYWVRANVTPEYQQ-E--QLAGYLSVRNTPNA-	49613026
19 Ypes_8-120	WSSSLVKNRRKNGDYYWVRANATPLRN-G--LTGYMSVRIAPTRA-	15979682
20 Ecar_8-121	IWTAVVKNRRKSGDYYWVKASTTPLMKE-G--KITGYMSVTRVRSQE	49611454
21 Ecar_8-118	IWTGIVKNRRKNGDHYWVKSSSTTPLRMKG-G--KITGYMSVRTA---	49611455
22 Reut_8-121	SWVGIVKNRRKNGDYYWVSATVPTPID-G--PLVGYTSVRSMATRE	53762136
23 Rmet_8-121	SWVGIVKNRRKNGDHYWVQATVPTPRVG-D--RVVGYTSVRSMASRE	48769261
24 Avin_10-121	KWNGIVKNRRKNGDHYWVRANVTPEYEG-D--RLVGYTSVTRASRR	23102254
25 Rsol_8-121	SWRGLVKNRRKDGCFYWVQANVTPVLQD-G--AIAGYTSVRSATPA	17430723
26 Bbro_8-121	SWLGVVKNRRKDGCFYWVLANATPIYED-G--EVVAYSSVRVKASDE	33575988
27 Tden_8-121	PWTGLVKNRRKNGDHYWVQANVTPVRED-G--VLSGFLSVRTPTRE	52006288
28 Bjap_8-121	PWLGAIVKNRRKNGDHYWVLATASPIREN-G--VTGYTSIRTRLPAD	27351190
29 Rpal_8-121	PVVGIVKNRRKNGDHYWVLASATPLWEN-G--QVTGYMSVTRKLPAD	39648595
30 Daro_8-121	PWTGIVKNRRKNGDHYWVLATVPTIREG-G--EILGYMSVRRKASAQ	53729524
31 Neur_4-121	PWSGMVKNRRKDGCFYWVYANVTPIREH-G--VTGHMSVRSKPTRD	30180852
32 Mfla_8-121	PWTGIVKNRRKNGDHYWVLANATPMRON-G--NVVGYMSVRRKPSRQ	53760232
33 Tden_8-121	PWSGLVKNRRKNGDYYWVDADVAPLIEN-G--KITVGYASIRSPDRE	52007875
34 Rgel_8-120	PWSAPVKNRRKDGCFYWVIANVTPLMSD-G--QPTGYMSVRTLPDR-	47574282
35 Bfun_4-120	PWTALVKNRRKNGDHYWVHANVTPEVEK-G--TVVGYLSVRVKPER-	48782623
36 Bbro_8-121	PWLAVVKNRRKSGGCFYWVLNAMPVIEA-G--VTGYASVRVKATQA	33576529
37 Lint_8-121	PWSGLIKNRRKSGDYYWVDATVTPVMNE-G--VISGYMSVRKKATED	24193388
38 Psyr_8-121	PWMGIVKNRRKNGDHYWVNAYVTPVLEN-R--VVGFSVRIKPTAE	23471893
39 Psyr_8-121	PWMGIVKNRRKNGDHYWVNAYVTPVLEN-R--QVVGFSVRIKPTAE	28852458
40 Pput_8-121	PWMGIVKNRRKSGCFYWVNAYVTPIFDN-N--VVGFSVRRVPTAE	24983638
41 Pflu_8-121	PWMGIVKNRRKNGDHYWVNAYVTPVFEEN-N--VVGFSVRRVPTAE	48731902
42 Paer_8-121	PWMGIVKNRRKNGDHYWVSAYVTPIIDQ-G--VVGFSVRRVPTAE	53727763
43 Psyr_8-122	PWMGVKNRRKSGGCFYWVSAYVTVAYEN-G--RIVGYESVRRLPTRD	46187912
44 Psyr_8-125	PWMGVKNRRKSGGCFYWVSAYVTVAYEN-G--RIVGYESVRRLPTRD	28852093
45 Pput_8-121	PWMGIVKNRRKNGDHYWVSAYVTAIYEQ-G--RISGYESVRRPTRE	24986255
46 Pput_8-120	SWMGIVKNRRKNGDHYWVNAYVTPILEG-G--RVVGYESVRRPCTR-	4545127
47 Pput_8-120	SWMGIVKNRRKNGDHYWVNAYVTPILEG-G--RVVGYESVRRPCTR-	24983799
48 Vvul_8-121	SWMGIVKNRRKNGDHYWVDASFPIREN-G--KVVEYQSVRRCPSR-	37200780
49 Vcho_8-120	SWMGIVKNRRKNGDHYWVDASFPIREN-G--KVVEYQSVRRCPSR-	9658429
50 Vpar_4-117	HWMGIVKNRRKNGDHYWVDASFPIREN-G--KVVEYQSVRRCPSR-	28808762
51 Fpro_1-115	SWMGLVKNRRKNGSYYWVSASFVPTINDKG--EVVEFQSVRRSPDQA	46916433
52 Vvul_5-119	SWMGLVKNRRKNGSYYWVSASFVPTINDKG--EVVEFQSVRRSPDQA	37201131
53 Sone_6-121	PWRGIVKNRRKNGDHYWVDAYVSPIMIN-G--QVVEFQSVRRSPDQA	24346080
54 Sthe_8-121	RWVGIVKNRRKNGDHYWVKAFVSPVLED-G--KIIIGFQSVRRKPTRE	51856248
55 Sone_14-127	SWRGLVKNRRKNGDHYWVDASFVSPIFEN-G--TIVGYQSVRRLPQA-	24347109
56 Ecol_6-121	SWMGLVKNRRKNGDHYWVSAYVTPIAKN-G--SIVEYQSVRRKPEPE	48195
57 Vcho_11-124	SWMGLVKNRRKNGDHYWVSAYVTPIAKN-G--SIVEYQSVRRKPEPE	9654939
58 Pflu_6-123	SWMGLVKNRRKNGDHYWVSAYVTPVTON-G--AAVEYQSVRRTPDTR	48730183
59 Vpar_8-121	AWRGLVKNRRKNGDHYWVDAYVTPIYEQ-N--QVVGYSQSVRRKPRE	28809550
60 Vvul_8-121	AWRGLVKNRRKNGDHYWVDAYVTPIYEN-G--VMSGYQSVRRKPRE	37201308
61 Vcho_24-136	AWRGLVKNRRKNGDHYWVDAYVTPIYEQ-G--QLTGYQSVRRKAER-	9658074
62 Ppro_6-119	AWRGLVKNRRKNGDHYWVDAYVTPIYEQ-G--KIIIGFQSVRRKPTQ	46915609
63 Ppro_8-121	PWRGIVKNRRKNGDHYWVDAYVTPIYON-D--RVSGYQSVRRPTPA	46915320
64 Sone_6-119	PWMGIVKNRRKNGSYYWVNAYVAPVYED-G--KIHEYQSVRRQATPE	24349621
65 Rpal_7-121	EIFPGYVKNRRKNGDHYWVFAHVTPEDEHG--NIVGYHNNRRTPDP-	39937542
66 Rpal_6-120	EIFAYVKNRRKNGDHYWVFAHVTPEDEHG--NIVGYHNNRRTPDP-	39937544
67 Rpal_7-121	EIFAYVKNRRKNGDHYWVFAHVTPEDEHG--NIVGYHNNRRTPDP-	39937263
68 Mmag_1-109	EIFAYVKNRRKNGDHYWVLAHVTPTEDEHG--KIVGYHNNRRTPPR-	46200832
69 Bjap_7-122	EIFAYVKNRRKNGDHYWVLAHVTPTEDEHG--KIVGYHNNRRTPPR-	27378086
70 Rpal_1-124	EIFAYVKNRRKNGDHYWVLAHVTPTEDEHG--KIVGYHNNRRTPPR-	39937372
71 Mmag_20-133	EIVFAYVINRRKNGDHYWVFAHVTPTVNTG--KIIIGFQSVRRAPSR-	46203509
72 Mmag_8-122	EIVFAYVINRRKNGDHYWVFAHVTPTVNTG--KIIIGFQSVRRAPSR-	23015159
73 Wsuc_9-123	EIFAYVKNRRKNGSYYWVLANTPSPFNKR--EIVGYHNNRRTPSPQ	34557986
74 Wsuc_7-121	EINAYVKNRRKNGSYYWVLANTPSPFNKR--EIVGYHNNRRTPSPQ	34557986
75 Wsuc_7-121	EIFAYVKNRRKNGSYYWVLANTPSPFNKR--EIVGYHNNRRTPSPQ	34557107
76 Daro_1-109	EIVFAYVINRRKNGDHYWVFAHVTPTVNTG--KIIIGFQSVRRAPSR-	46140878
77 Cjej_1-113	EIFAYVKNRRKNGDHYWVFAHVTPTVNTG--KIIIGFQSVRRAPSR-	15792513
78 Cvio_9-125	EIVFAYVINRRKNGDHYWVFAHVTPTVNTG--KIIIGFQSVRRAPSR-	34495851

PAS_Che

79 Ypes_15-131	PISDNIKRIKNGDHYWVLQGTYPVVDQGG--NVIEIKIASDVTER	21958470
80 Ypes_1-112	PISDNIKRIKNGDHYWVLQGTYPVVDQGG--NVIEIKIASDVTER	15980509
81 Ypes_11-127	PISDNIKRIKNGDHYWVLQGTYPVVDQGG--NVIEIKIASDVTER	51590152
82 Paer_17-134	PORGVFKRLRDGCFYVWLEATYFPVKNAEG--AVVEVLKIAADVTRN	53727981
83 Paer_17-134	PORGVFKRLRDGCFYVWLEATYFPVKNAEG--AVVEVLKIAADVTRN	9947925
84 Vpar_13-130	HKRGTFERILKSGCFYVWLEATYFPVKNAEG--AVVEVLKIAADVTRN	28808720
85 Vcho_53-169	SHSGTFMRKKKDGSLVWLEATYFPVLD--D-G--VSSVMKIASDVTEQ	9658296
86 Vcho_23-140	AQRKMFHRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	9655902
87 Pput_23-138	AISGTFERILKSGCFYVWLEATYFPVKNAEG--AVVEVLKIAADVTRN	24985060
88 Pflu_23-138	PISGTFLLRLKSGCFYVWLEATYFPVKNAEG--AVVEVLKIAADVTRN	48730737
89 Agam_1-111	VKSGLIFSRILKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	31195287
90 Pput_148-265	YIAERFKRILKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	24982183
91 Pflu_140-257	FVAGRFKRVDSHGFYVWLEATYFPVKNAEG--AVVEVLKIAADVTRN	48730078
92 Psyr_145-261	FVAGRFKRVDSHGFYVWLEATYFPVKNAEG--AVVEVLKIAADVTRN	46188225
93 Psyr_170-286	FVAGRFKRVDSHGFYVWLEATYFPVKNAEG--AVVEVLKIAADVTRN	28851464
94 Psyr_140-257	FIVDRFRRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	46188835
95 Psyr_155-272	FIVDRFRRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	28853310
96 Vpar_141-258	FIVDRFRRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	28808847
97 Psyr_125-242	FVAGRFKRVDSHGFYVWLEATYFPVKNAEG--AVVEVLKIAADVTRN	28854157
98 Psyr_140-257	YVEGRFRRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	28855155
99 Pput_143-260	YHSRFRRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	24985060
100 Pflu_143-260	YHSRFRRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	48730737
101 Psyr_144-257	FIGDQFKRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	23471312
102 Psyr_143-257	FIGDQFKRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	28855718
103 Paer_140-257	YVVGQFRVRHNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	53727683
104 Paer_117-234	YVVGQFRVRHNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	9947371
105 Pflu_144-257	YFQGGFRRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	48730703
106 Bjap_134-256	YFQGGFRRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	27349173
107 Rpal_135-250	YFQGGFRRLKNGDHYWVLEATYFPVKNAEG--AVVEVLKIAADVTRN	39647354

Figure C.2 continued

108 Bja _p _263-378	YQAAEYKRIKGGG:EVYIQASYNPILDLNG--KPFKVVKYATDITKQ	27349173
109 Rpa _l _256-371	YQAAEYKRIKGGG:EVYIQASYNPILDLNG--RPFKVVKFATDITKQ	39648804
110 Rpa _l _255-372	YQAGEYKRIKGGG:EVWIQASYNPILDLNG--RPFKVVKYAADITRQ	39647354
111 Rpa _l _133-249	YQAAEYKRIKGGG:EIWLQASYNPIFDKGG--RPAKVVKFATDVTQE	39648804
112 Bsp_ ₉₀₋₂₀₅	FQSAQYKRIKNGG:IVWIQASYNPILDLNG--KPKVVKVFATDISAQ	18033717
113 Rpa _l _11-128	YQAGEFHRIKGGG:EVWIQASYNPILDKNG--KPTGVVKFAADITAA	39647354
114 Bja _p _19-134	QQAEEFKRIAKGGG:EVWIEASYNPVPDNAG--KPKVVKVIATDITAK	27349173
115 Rrub_ ₉₀₋₂₀₅	QQAQFMRIKGGG:VWVLEASYNPILNLDG--KPKVVKVFATDISGR	48764139
116 Mma _g _103-220	YQSGEYKRIKGGG:EVWYIGSYNPVPDSEG--KLYAVVKFANDITEA	46203332
117 Xcam_ ₁₃₈₋₂₅₅	FDAGRYKRVGRDGG:EVWYIQASYNPVLDERG--RPYKVIKYATDITRQ	21112820
118 Xcit_ ₁₆₁₋₂₇₈	FDAGRYRRIKDDGG:EVWYIQASYNPVLDEHG--RPYKVVKYATDITRQ	21107943
119 Xcam_ ₂₆₂₋₃₇₇	FDAGRYRRLRKDDG:TAWYIQASYNPILDVSG--RPYKVVKYATDITDQ	21112820
120 Xcit_ ₂₈₅₋₄₀₀	FDAGRYRRLKDDG:AVWYIQASYNPILDVSG--RPYKVVKYATDVTDQ	21107943
121 Xcam_ ₁₈₋₁₃₃	FHAGRYCRLNGGE:EVWYINASYNPILLDRSG--KPYRVVKYATDITAQ	21112820
122 Xcit_ ₄₁₋₁₅₆	FNAGRYCRLKRGGE:EVWYINASYNPILLDRSG--KAYRVVKYATDITAQ	21107943
123 Atum_ ₈₋₁₂₄	FDQQQYKRIKGGG:EVWIEASYNPVMR-RG--KPKVVKVIATDITAQ	6498287
124 Atum_ ₁₃₋₁₂₉	FDQQQYKRIKGGG:EVWIEASYNPVMR-RG--KPKVVKVIATDITAQ	15163624
125 Atum_ ₁₀₋₁₂₄	FDQQQYKRIKGGG:EVWIEASYNPVFR-RG--KPKVVKVIATDITER	15160314
126 Atum_ ₉₋₁₂₅	YDQGGYRRQAKNGG:EIWIEASYNPVFR-FG--KPKVVKVIATDITVI	15160002
127 Sent_ ₈₋₁₂₄	LDSNSYRRLAKGGG:EIWYIQASYNPVFR-NG--KPKVVKVFATDITAA	11545452
128 Atum_ ₈₋₁₂₄	FDAREYRRIKDDG:AVWIEASYNPVFR-SG--KPKVVKVIATDITAK	17741118
129 Ccre_ ₄₃₋₁₅₆	FDAREYRLRKDDGG:EVWYIQASYNPVKNASG--KVTRIKLVATDITAQ	13425051
130 Ssp_ ₂₁₋₁₃₈	PASGEFHRIKGGG:DVWIDAHYTPILDAAG--KPKYKVKLATDITAQ	52010400
131 Psyr_ ₁₄₁₋₂₅₆	YDSGEYKRIKNGG:ELWISATYNPIFDPDG--RPYKVVKFANDVTES	23468713
132 Psyr_ ₁₄₁₋₂₅₆	YDSGEYKRVKNGG:ELWISATYNPIFDPDG--RPYKVVKFANDVTES	28852878
133 Psyr_ ₂₆₃₋₃₇₈	YDANEYKRRKDDGG:EIWYIQATYNPIFDAQG--KPKYKVKFALDVTVA	23468713
134 Psyr_ ₂₆₃₋₃₇₈	YDANEYKRRKDDG:EIWYIQATYNPIFDAQG--KPKYKVKFALDITEA	28852878
135 Psyr_ ₁₈₋₁₃₄	FDEGQYKRLKNGG:EIWLQATYNPVPDEQGG--NPFKVVKFATDVTQA	23468713
136 Psyr_ ₂₁₋₁₃₄	FDEGQYKRLKNGG:EIWLQATYNPVPDEQGG--NPFKVVKFATDVTQQ	28852878
137 Naro_ ₂₄₂₋₃₅₇	YVSGEYKRLTKAGG:EIWIRASYNPIIGTDG--KPTKICKIAADVTQA	48847895
138 Vcho_ ₁₈₄₋₂₉₉	FVDEYKRFKGGG:EIWYIQASYNPIMDSEG--KPKYKVKYATNTVQR	9654496
139 Ccre_ ₂₀₋₁₃₅	FLTGKYKRVKGGG:EIWYIQATYNPVPDRHG--KLAKVVKFAADITAA	13423061
140 Bba _c _38-159	FVAGEFKRIAKGGG:EIWYINASYNPIEDNEG--NPKVVKVFATDITAV	39576448
141 Bba _c _165-281	FEAGEFKRFKAGG:EVWYINASYNPIEDNAG--KPKVVKVFATDITAD	39576448
142 Ssp_ ₂₆₄₋₃₈₁	TRTGEFRRLKDDG:ELWYIQATYNPIFDGNG--DVKVVKFPAVDITDQ	52010400
143 Atum_ ₁₃₁₋₂₄₆	LMADEFMRLKGGG:KVFYIQASYNPIEDMNG--RVFKVVKFATDVTTR	6498287
144 Atum_ ₁₃₁₋₂₄₆	LVADEFMRLKGGG:KVFYIQASYNPIEDMNG--KVFVKVVFATDVTGR	15160314
145 Atum_ ₁₃₂₋₂₄₇	FSTGQFMRLKDDG:RVFYIQASYNPIIDDRG--RVFKVVKFPAVDITDR	15160002
146 Sme _l _131-246	FIANEYVRYKGGG:EVWYIQAYNPVIRDPNG--RVYKVVKFATDITER	11545452
147 Atum_ ₁₃₁₋₂₄₆	FISDEFVRYKNGG:EIWYIQAYNPVILDEAG--KVVVKVVFATDVTFR	17741118
148 Ccre_ ₁₆₃₋₂₇₈	FVAAEFRRKGGG:EVWYIQASYNPVPDAGK--RVVKVVKFATDVTGR	13425051
149 Ssp_ ₅₀₉₋₆₂₆	SDTGKYQRFKGGG:EIWLSASYSPIFDQEN--NVVKVVKFANDITTA	52010400
150 Psyr_ ₃₈₅₋₅₀₀	FVSGRFMRVKYGG:KIWIQATYSPIFDHGG--LFFKVVKFATDITRQ	23468713
151 Psyr_ ₃₈₅₋₅₀₀	FHSGRFMRVKYGG:KIWIQATYSPIFDHGG--LFFKVVKFATDITRQ	28852878
152 Bba _c _288-403	FDSGRYLRVKAGG:SIWYIQATYNPIMDMNG--KPKYKVKFASDITQQ	39576448
153 Naro_ ₃₆₄₋₄₇₉	FHAGRFHRVKYDR:DVWYIQATYNPIFDLRG--KPVVKVVFATDITDQ	48847895
154 Crio_ ₁₈₋₁₃₅	IKDQFRRIRKDDG:TVWLEASYNPVPDKDGG--QVVKVMVFALDVTQE	34102765
155 Cje _j _24-137	ARSGLFRRIRKGGI:DVYLEANYLPISDNGG--YVYKIKFANDITQR	6968544
156 Ssp_ ₁₈₋₁₃₂	PLTDQFPRIRKDDGG:VWYIQATYAPFFNEDG--TVDRIVKIASNITFR	52011981
157 Mdeg_ ₁₇₋₁₃₄	SLSGEIERVKNNG:SVLLHATYSPLLEDGG--SVYRVVKFATDITEV	48863724
158 Ccre_ ₁₋₁₀₃	PIITKTQLFKGGG:IVRVRAAYSVPVLDTAG--KLKRVILFATDVSEL	13424414
159 Cje _j _142-259	FQSGKYIRYKRNK:KVYLEASYNPVPKNDGG--KIYKVKIKFATDISEQ	6968544
160 Agam_ ₅₁₋₁₆₆	FKSGRFRRFKQGG:EIWLEATYNNAVGRSHG--KVKIKVKFASNITSQ	31194935
161 Agam_ ₁₁₆₋₂₃₁	FKSGLFLRRNSHG:AIWLEATYNPICDESG--KVNVRVKFASDITER	31195287
162 Vcho_ ₁₄₅₋₂₆₀	FKSGLFHRRLRHG:DLWLEATYNPIFNADG--VVTQVVKFASDITDQ	9655902
163 Vpar_ ₁₃₅₋₂₅₀	FKSGQFLRRSASGG:KVYIEATYNPIFDPDG--NVIKVVKFASDITDK	28808720
164 Vcho_ ₁₇₄₋₂₈₉	AYSGRFLRRNSYG:QVWYIQASYSVPVKDQNN--KVKYKVKFASDITEQ	9658296
165 Paer_ ₁₃₉₋₂₅₄	FSRGHFERRKAGG:RVVHIEATYNPVPKDGSG--RIIKVKIKFALDITEQ	53727981
166 Paer_ ₁₃₉₋₂₅₄	FSRGHFERRKAGG:RVVHIEATYNPVPKDGSG--RIIKVKIKFALDITEQ	9947925
167 Paer_ ₂₀₋₁₃₅	HFSGRCKRITREG:PLWLEATYNPVPDQGG--RLLKVVVKYASDIDAI	53727683
168 Paer_ ₃₋₁₁₂	HFSGRCKRITREG:PLWLEATYNPVPDQGG--RLLKVVVKYASDIDAI	9947371
169 Crio_ ₁₄₂₋₂₅₉	FLEGQFRMNSHGG:EVWYIQATYNPVLDPGG--RPRVVKVVFASDISEQ	34103010
170 Ypes_ ₁₃₆₋₂₅₃	FIIGRFERLNRRGG:RVWLEASYNPIMDNEG--NVLKVVVKIAQDITAI	21958470
171 Ypes_ ₁₃₂₋₂₄₉	FIIGRFERLNRRGG:RVWLEASYNPIMDNEG--NVLKVVVKIAQDITAI	51590152
172 Psyr_ ₂₀₋₁₃₅	FISGTFKRINKNGG:SVWLEASYNPVIDTQGG--KVVVKVVKYALDVTRR	28855718
173 Psyr_ ₂₀₋₁₃₅	FISGTFKRINKNGG:SVWLEASYNPVIDTQGG--KVVVKVVKYALDVTRR	23471312
174 Pflu_ ₁₈₋₁₃₅	FQSGTIFERRKSGG:PIWLEASYNPIKDDSG--RVVKVVKYAMVDITAK	48730703
175 Crio_ ₂₀₋₁₃₇	FQAGHFLRRRKDDG:DIWLEASYSPILDGDD--RVTVGVVKVATDITQQ	34103010
176 Crio_ ₁₄₀₋₂₅₇	FVSGRFERLHSGG:PIWLEASYNPIFDGDD--VVFVKVKFATDITDQ	34102765
177 Lmon_ ₈₋₁₂₄	KFQNKIERKNARGG:RVWFEATYIPIIRED--TVGVVKVAKIATDITRR	47013964
178 Lmon_ ₈₋₁₂₄	KFQNKIERKNARGG:RVWFEATYIPIIRED--TVGVVKVAKIATDITRR	46881199
179 Linn_ ₈₋₁₂₄	KFQNKIERKNARGG:RVWFEATYIPIIRED--TVGVVKVAKIATDITRR	16414310
180 Esp_ ₁₄₋₁₃₀	TYQDKIERRSADGG:QKWLEATYMPIFEDG--RRVGVGVSKIATDITVR	46113970
181 Oihe_ ₉₋₁₂₆	SYQNKIKRKDANG:PIWLEATYMPVFEEDN--KVAISKIATNITER	22776939
182 Esp_ ₁₁₋₁₂₇	SSADKIERIDARGG:SIWLEATYMPIYEAN--KVGGVVKIASDITDR	46113993

Generic PAS

183 Psyr_ ₂₀₋₁₃₅	HFSGTLRLVSKTGS:RVWLRITIVVPYKNETG--GVEQITLYSSVLTRT	46188835
184 Psyr_ ₂₀₋₁₃₅	HFSGTLRLVSKTGS:RVWLRITIVVPYKNETG--GVEQITLYSSVLTRT	15077778
185 Psyr_ ₃₅₋₁₅₀	HFSGTLRLVSKTGA:RVWLRITIVVPYKNETG--GLEQITLYSSVLTRT	28853310
186 Pput_ ₂₆₋₁₄₃	HLNGAFRLKNGG:EAWLRISILQPVKNSEG--RIKYFTLHSSDLTRT	24982183
187 Pflu_ ₁₈₋₁₃₅	HFAGAVRLSRNGG:EAWLRISIVQPIRSSDG--RIKHFSIFSSDLTRT	48730078
188 Vpar_ ₁₉₋₁₃₆	FWNGAQITKGGG:EAWLRVVIQPVNRASD--QVESFFVFASDLTRT	28808847
189 Psyr_ ₂₂₋₁₃₉	SVIDRYRFLHADGRL:LIWIRALWQPVLDQGG--RLMTLCQYGSIDITQI	46188225
190 Psyr_ ₄₇₋₁₆₄	SVIDNYRFLHADG:LVWVRAMWQPVLDQGG--KLVTLCQYGSIDITQI	28851464
191 Psyr_ ₁₇₋₁₃₅	SMTNEYHFIADNNT:IACLKLFWFPVQAENG--GLSHIQCYGHDVTKS	28855155
192 Vcho_ ₆₃₋₁₇₇	-PVSAELQLRSQGE:AIWIKADLYPIKAING--QLQNVVVLQDITAA	9654496
193 Ssp_ ₃₈₉₋₅₀₄	PFYDEILNNTKDDGE:AYWISLAINPVPDEAG--KLQKFVSIQTNITDV	52010400
194 Psyr_ ₈₉₋₂₀₉	PYRIKNRLAMKDGTYRWYFAQGETLRDARG--TPLRVAGSLRIDHDE	46188184
195 Psyr_ ₈₈₋₂₀₉	PYRIKNRLAMKNGTYRWYFAQGETLRDARG--TPLRVAGSLRIDHDE	28854917
196 Crio_ ₉₂₋₂₁₃	PYDIEYRLQCKNGDYRWFRFARGATLRDGG--VPLRVAGSLRIDITAI	34103227
197 Psyr_ ₂₃₁₋₃₄₈	PFDIEYRLKMKTGEXRWFRFARGQTRRNPEG--VPLRVVGVGVVHLK	46188184
198 Psyr_ ₂₂₇₋₃₄₈	PFDIEYRLKMKTGEXRWFRFARGQTRRTPEG--VPLRVVGVGVVHLK	28854917
199 Crio_ ₂₃₁₋₃₅₂	PYDLDYRLQCKNGEXRWFRFARGQTRRAADG--APLRVVGVGVVHLK	34103227
200 Rrub_ ₄₁₋₁₆₂	GVDVTRYRLKMRDGA:YRWFRATGGCLRDAG--KPLRACGSLTDVHDE	48766385
201 Rrub_ ₁₇₃₋₂₉₄	GVDVTRYRCKVRNGS:YRWFRATGGCLRDAG--NPLRACGSLINVDVA	48766385
202 Rpa _l _23-144	FYDVYRRLVKDGS:YRWFRATGGGVVLDENR--KPRRACGSLVIDIEL	39648851
203 Rrub_ ₂₈₋₁₄₉	SYDVYRRLKRDGA:YHWFRAMGGVNRDAG--KATRMCGSLVIDIAE	48764228
204 Mma _g _29-147	PYQVEYRIKADGS:TVFIOEQAHFVNDKGG--NLAVVDGVFVLDITQQ	20904696

Figure C.2 continued

205 Mace_31-149	AYQVEYRIKKT DGS	TVFIQELAHLVND DAG--NLAYIDGVFLDVT PQ	19917084
206 Mmaz_29-147	PYQVEYRIQKSCGD	TVFVQEQARLVNDEHG--NIAYIDGVFLDVT PQ	20906162
207 Mbur_32-150	PKVNYRIKTKNGD	TVYIREEGKLVNDEHG--NAAYLDGVFLNITEN	46142191
208 Mmaz_153-269	FNFERALKLKAQDK	PLHTVTSSVPIKDDTG--AIVANLTIIITDMTEM	20904696
209 Mace_155-271	YNLERPLKLRALDK	PLHTVTSSVPIKDDSG--AIIGNLTIIITDMTEM	19917084
210 Mmaz_153-269	YNLEKSIKFKALDK	PLFTVLSAVPVKDDTG--TIAGSLMVTDMTEM	20906162
211 Mbur_156-284	ESLEAIIKLHGLD	ELYTISSASPIFDEEG--VFEGILEVLTDLTDI	46142191
212 Aful_169-283	VINHQA VTRTKDGR	EVPLVLSCLIPVY--VDG--EMVGVLDLFTDITEL	2649560
213 Aful_189-305	IEHNEVVKLV-KDGS	EFIASASIPVYVY--VDG--EFAGYIEVYFIDITEL	2649548
214 Aful_62-177	IEGKEGFEVVKTG	KAMPILT-CA-VY--VDG--EFGMVDFIDITEL	2649548
215 Aful_35-157	KIENQEVKLGTEK	LMHILYTSAPVK--VNG--ELVGMGVFYVDVTPI	2649560
216 Hsp_8-123	IR-TSVRTSELPA	QAHAKASATPLH--NDG--AVIGAVEVLTIVTDV	10580941
217 Aful_319-440	YLTEIWLNFERNKG	KAYVRATAAPVYNKGG--ELIGVVESTIEDITEI	2649548
218 Ddes_384-499	PVKADIMSIDRDG	CTLYVKPSADVLRDAQG--RKMGYVEVASVITDL	53691249
219 Dvul_426-541	RYEAEIIEIINHGS	SRWIRPYGDILHDCG--QKAGYLEVASDVTEI	46448762
220 Dpsy_501-615	IITDNTIARPDQGI	IIPIKYTGAPIKDAGK--NIKGALEYILDVTEE	50877249
221 Dpsy_379-493	IITDNTIARPDQGI	IIPIKYTGAPIKDAGK--NIKGALEYILDVTEE	50877249
222 Dpsy_257-371	VITDHTIARPDQG	IIPIKYTGAPIKDAGK--NIKGALEYILDMTEE	50877249
223 Dpsy_135-249	VITDHTIARPDAG	IIPIKYTGAPIKDAGK--NIKGALEYILDMTEE	50877249
224 Dpsy_623-736	VVDRTTASP-NG	EMSIKYTGSPDKAGK--NIKGALEYITDVTEI	50877249
225 Dvul_393-508	RRNPFSTYSTVGE	QFEMIDVMPIRDAGS--DIIGGITFWNDVTEI	46448995
226 Dvul_261-373	QCGN-ISYRRDDG	VFPLHYEVSPILDRDG--AVNGAIAVIDLTFE	46451007
227 Ddes_360-475	ISNLEVTITGHKGR	OTEVLANVTYLDQMEG--TITGGMCLYLDMTQE	53691138
228 Dvul_359-474	ITNIDVTIK-HKGG	EVQVLANVTYLDATDG--NIFGGFCLYLDVTEA	46449695
229 Dvul_118-232	IARREVDLVKRGK	KRRRLHINASPLYDLG--TLMGALCIYQDLTEL	46448495
230 Ddes_118-232	VTGREVDLTKRGN	TRRIQIHASPLFDLGG--GLMGALCIYQDLTEL	53691581
231 Dvul_384-499	LRNIEVDLTKSGE	TVHTLVDAVPLSDLDG--SLIGSTIYADITEL	46450563
232 Ddes_263-378	ITGVEVEFT-RDGE	TRYSVLDSSPIYDLG--NLMGAFVCTDMTEF	23473878
233 Dvul_366-480	LHDVA-VWNAPSGR	EVHLNVAATPFYDMGD--ELLGSIAFWMDITDI	46450119
234 Dvul_389-504	VEAIEREMRDTSGH	VRRVRLDAAPLNDLDG--SLIGALIALADITVI	46449710
235 Ddes_379-494	VTI-VEREMRTEKGN	MRHVRIDAAPLNDLDG--RIGALIALADITDI	53691131
236 Dvul_405-518	-TQTEENLR-KKGD	LRTARLTAAPLHORDG--NIGATVLELDL	46448526
237 Xcam_140-257	SKAVRAPE----	FG-ROIDFVYSPILAADG--TKLGTIAQWMDVTAQ	21114304
238 Xcit_156-271	SKAVRAPE----	FG-ROIDFVYSPILAADG--TKLGTIAQWMDVTAQ	21109549
239 Xcam_163-285	TVSERYTF----	GP-VTLDLTMTPLYAGDG--RRSGMMLWRNVSAE	21112986
240 Xcit_164-284	TSERYRFE----	GP-VTLDLTMTPLYAPDG--RRSGMMLWRNISAE	21108110
241 Xcam_145-249	---GABAL-----	GA-LQLQES-TVVHDDG--TFLGVVVEWDRQS	21111282
242 Xcam_277-395	TTTFEERF-----	G-AVFAQTVTTIQ-DEGD--QNVGDVCEWRDRTIE	21112973
243 Neur_277-373	-----	RTYSLLLMPVT-ESG--ERA-AVVEWDRDTE	30180852
244 Vvul_42-160	PYSTVIAI----	DD-VRLNLNVGAMLDAG--NYVGNLTLEWQDVTE-	37201894
245 Vvul_38-156	PYSTVIAI----	DD-VRLNLNVGAMLDAG--NYVGNLTLEWQDVTE-	27359118
246 Vcho_38-157	PYTTVINI----	KG-VKTELIVGAILDRG--SYIGNTLEWRDVTEE	9658538
247 Vvul_169-288	PYRTDIIV-----	GD-IRIELNVAAVKNAGK--EYIGNSLEWRDVTEQ	37201894
248 Vvul_165-284	PYRTDIIV-----	GD-IRIELNVAAVKNAGK--EYIGNSLEWRDVTEQ	27359118
249 Ssp_146-263	PFKTDITV-----	GE-MKFALNVDGIFDDGD--TYVGNLLEWADVTEA	52010400
250 Vvul_296-413	PFSSDIKV-----	GS-LEFNLTCIAMRDGSG--NYMGPALQWIDITEQ	37201894
251 Vcho_164-282	PFSTMIKV-----	GS-LEFNLTCIAMRDTKG--EYIGPALQWIDITEQ	9658538
252 Rrub_9-120	PHHAVIAL-----	GD-EFLDLQIEALGE-RA--APKAVLTWSIVTER	48765139
253 Rrub_130-241	PHSKIRL-----	GE-EFLDLRVTAIFGERG--EYEAMLLCWSVSTHL	48765139
254 Mmag_261-371	PHVANISL-----	GQ-EVIELNVSAAILDRKG--HYVGPLLTWMPVTEK	46201816
255 Mdeg_138-254	PKTTTLNI-----	GD-MVFLIATPWFNSNG--ERLGTLEWLDKTEE	48863724
256 Rpal_13-128	VHRATIQT-----	GI-RIFDLIATPINNADG--SRAGVVVEWADASIR	39648804
257 Sone_145-261	KYESQIQV-----	AS-CHPFLTASPIILTSG--ERLGSVVEWLDRTTE	24348039
258 Mdeg_527-640	TYSTQIKV-----	GI-RTFSLIANPITKD-G--ERVGTVVEWLDRTTE	48863731
259 Mdeg_269-384	TYNGGAKV-----	GG-RSFTLVIANPIF--VDG--KKIGAVVEWLDRTAE	48863724
260 Cvio_457-573	T-RAEIQV-----	AG-RTFSLVANPVFADAG--ERLGSVVEWLDRTAE	34332882
261 Paer_173-289	V-KAELNL-----	GG-RRFSLDVPVFENDAN--ERLGSVAVQWLDRTTE	53726991
262 Paer_173-289	V-KAELNL-----	GG-RRFSLDVPVFENDAN--ERLGSVAVQWLDRTTE	9946009
263 Daro_436-552	THRATIRL-----	GG-RVFALTVPVFNTRG--GRLGFVVEWLDRTNE	53729524
264 Sone_21-137	SHTAQISI-----	GK-RIFKLILTPILSRDN--KHLGTGVVEWLDRTES	24348039
265 Daro_261-377	IHRTEIDI-----	GG-RYFSLVACPIVNDQG--ERHGTVVEWLDRTAE	53729525
266 Cvio_181-297	S-RSSISV-----	GG-RTFGLILTPILKAGK--ERLGDVVEWQDNTTE	34332882
267 Cvio_319-435	THRSSIVV-----	GG-RTFGLILSPIFNDKG--DRLGAVVEWQDNTTE	34332882
268 Cvio_43-159	TRRGTIKV-----	GG-RTFSLVLTPIRDQGN--RKLGA VVEWLDRTAE	34332882
269 Lint_219-335	TRRSSITI-----	GG-REFDLIANPIVDVNG--NKLGTVVEWSDVTEQ	24196190
270 Lint_219-335	TRRSSITI-----	GG-REFDLIANPIVDVNG--NKLGTVVEWSDVTEQ	45600635
271 Lint_349-466	TRRSSINI-----	GG-RTFNLIANPIIDETG--DRLGSVVEWSDVTEQ	24196190
272 Lint_481-597	IHKATIKI-----	GG-RTFDLIANPIILDSNG--KRLGSVVEWSDVTNE	24196190
273 Cvio_19-144	QHTAMIFV-----	GE-VMFETKVFPIWDSANFSQLLCFMASFDVSSE	34103223
274 Gsul_12-134	PHSAEIPF-----	GG-ITLRTTSFPIWKKNPGRVKCYMACWDITAE	39985192

Figure C.2 continued

APPENDIX D

PUBLICATIONS

1. Shu, C.J., Ulrich, L.E., and Zhulin, I.B. (2003) The NIT domain: a predicted nitrate responsive module in bacterial sensory receptors.
2. Ulrich, L.E., Koonin, E.V., and Zhulin, I.B. (2005) One-component systems dominate signal transduction in prokaryotes, *Trends Microbiol*, **13**, 52-56.
3. Ulrich, L.E., and Zhulin, I.B. (2005) Four-Helix Bundle: a Ubiquitous Sensory Module in Prokaryotic Signal Transduction, *Bioinformatics*, in press.
4. Wu, M., Ren, Q., Durkin, A.S., Daugherty, S.C., Brinkac, L.M., Dodson, R.J., Madupu, R., Sullivan, S.A., Kolonay, J.F., Nelson, W., Tallon, L.J., Jones, K.M., Ulrich, L.E., Gonzalez, J.M., Zhulin, I.B., Robb, F.T. and Eisen, J.A. (2005) Life in Hot Carbon Monoxide: the Complete Genome Sequence of Carboxydotherrmus hydrogenofomans Z-2901, *PLoS Genetics*, preprint.
5. Richardson, P., Chain, P., Vergez, L., Malfatti, S., Larimer, F., Hauser, L., Land, M., Lao, V., Denef, V., Konstantinidis, K., Parnell, J., Sul, W., Tsoi, T., Marx, C., Smith, D., Chai, B., Ulrich, L.E., Zhulin, I.B., Latorre Reyes, V., Agulló, L., Gómez, L., Córdova, M., González, M., Seeger, M., Cook, A.M., LiPuma, J. (2005) The complete genome sequence of *Burkholderia xenovorans* LB400, a potent degrader of polychlorinated biphenyls, *PNAS*, submitted.
6. Ulrich, L.E., Black, W., Ma, Q., Taylor, B.L., and Zhulin, I.B. Resolving the Function of Chemotaxis PAS Domains through Protein Sequence Analysis, manuscript in preparation.

REFERENCES

- Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) Issues in Searching Molecular Sequence Databases, *Nat Genet*, **6**, 119-129.
- Altschul, S.F. and Gish, W. (1996) Local alignment statistics, *Method Enzymol*, **266**, 460-480.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic Local Alignment Search Tool, *J Mol Biol*, **215**, 403-410.
- Altschul, S.F. and Koonin, E.V. (1998) Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases, *Trends Biochem Sci*, **23**, 444-447.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.
- Anantharaman, V. and Aravind, L. (2000) Cache - a signaling domain common to animal Ca²⁺ channel subunits and a class of prokaryotic chemotaxis receptors, *Trends Biochem Sci*, **25**, 535-537.
- Anantharaman, V. and Aravind, L. (2001) The CHASE domain: a predicted ligand-binding module in plant cytokinin receptors and other eukaryotic and bacterial receptors, *Trends Biochem Sci*, **26**, 579-582.
- Anantharaman, V., Koonin, E.V. and Aravind, L. (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains, *J Mol Biol*, **307**, 1271-1292.
- Antoine, R., Huvent, I., Chemlal, K., Deray, I., Raze, D., Locht, C. and Jacob-Dubuisson, F. (2005) The periplasmic binding protein of a tripartite tricarboxylate transporter is involved in signal transduction, *J Mol Biol*, **351**, 799-809.
- Appleby, J.L., Parkinson, J.S. and Bourret, R.B. (1996) Signal transduction via the multi-step phosphorelay: Not necessarily a road less traveled, *Cell*, **86**, 845-848.

- Aravind, L. and Koonin, E.V. (1998) The HD domain defines a new superfamily of metal-dependent phosphohydrolases, *Trends Biochem Sci*, **23**, 469-472.
- Aravind, L. and Ponting, C.P. (1997) The GAF domain: an evolutionary link between diverse phototransducing proteins, *Trends Biochem Sci*, **22**, 458-459.
- Aravind, L. and Ponting, C.P. (1999) The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins, *Fems Microbiol Lett*, **176**, 111-116.
- Armitage, J.P. (1999) Bacterial tactic responses, *Adv Microb Physiol*, **41**, 229-289.
- Balazsi, G., Barabasi, A.L. and Oltvai, Z.N. (2005) Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli, *P Natl Acad Sci USA*, **102**, 7841-7846.
- Batchelor, E., Walthers, D., Kenney, L.J. and Goulian, M. (2005) The Escherichia coli CpxA-CpxR envelope stress response system regulates expression of the Porins OmpF and OmpC, *J Bacteriol*, **187**, 5723-5731.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) The Pfam protein families database, *Nucleic Acids Res*, **32**, D138-D141.
- Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0, *J Mol Biol*, **340**, 783-795.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2005) GenBank, *Nucleic Acids Res*, **33**, D34-D38.
- Bernal, A., Ear, U. and Kyrpides, N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide, *Nucleic Acids Res*, **29**, 126-127.
- Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions, *Nucleic Acids Res*, **29**, 2607-2618.

- Biswas, I. and Scott, J.R. (2003) Identification of rocA, a positive regulator of covR expression in the group A streptococcus, *J Bacteriol*, **185**, 3081-3090.
- Blue, C.E. and Mitchell, T.J. (2003) Contribution of a response regulator to the virulence of *Streptococcus pneumoniae* is strain dependent, *Infect Immun*, **71**, 4405-4413.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res*, **31**, 365-370.
- Bork, P., Holm, L. and Sander, C. (1994) The Immunoglobulin Fold - Structural Classification, Sequence Patterns and Common Core, *J Mol Biol*, **242**, 309-320.
- Boukhvalova, M., VanBruggen, R. and Stewart, R.C. (2002) CheA kinase and chemoreceptor interaction surfaces on CheW, *J Biol Chem*, **277**, 23596-23603.
- Bowie, J.U., Pakula, A.A. and Simon, M.I. (1995) The 3-Dimensional Structure of the Aspartate Receptor from *Escherichia-Coli*, *Acta Crystallogr D*, **51**, 145-154.
- Bru, C., Courcelle, E., Carre, S., Beausse, Y., Dalmar, S. and Kahn, D. (2005) The ProDom database of protein domain families: more emphasis on 3D, *Nucleic Acids Res*, **33**, D212-D215.
- Cai, S.J. and Inouye, M. (2002) EnvZ-OmpR interaction and osmoregulation in *Escherichia coli*, *J Biol Chem*, **277**, 24155-24161.
- Calogero, S., Gardan, R., Glaser, P., Schweizer, J., Rapoport, G. and Debarbouille, M. (1994) RocR, a Novel Regulatory Protein Controlling Arginine Utilization in *Bacillus-Subtilis*, Belongs to the Ntrc/Nifa Family of Transcriptional Activators, *J Bacteriol*, **176**, 1234-1241.
- Chai, W.H. and Stewart, V. (1998) NasR, a novel RNA-binding protein, mediates nitrate-responsive transcription antitermination of the *Klebsiella oxytoca* M5al nasF operon leader in vitro, *J Mol Biol*, **283**, 339-351.
- Chan, C., Paul, R., Samoray, D., Amiot, N.C., Giese, B., Jenal, U. and Schirmer, T. (2004) Structural basis of activity and allosteric control of diguanylate cyclase, *P Natl Acad Sci USA*, **101**, 17084-17089.

- Changela, A., Chen, K., Xue, Y., Holschen, J., Outten, C.E., O'Halloran, T.V. and Mondragon, A. (2003) Molecular basis of metal-ion selectivity and zeptomolar sensitivity by CueR, *Science*, **301**, 1383-1387.
- Cho, H. and Winans, S.C. (2005) VirA and VirG activate the Ti plasmid repABC operon, elevating plasmid copy number in response to wound-released chemical signals, *Proc Natl Acad Sci U S A*.
- Clamp, M., Cuff, J., Searle, S.M. and Barton, G.J. (2004) The Jalview Java alignment editor, *Bioinformatics*, **20**, 426-427.
- Comenge, Y., Quintiliani, R., Li, L., Dubost, L., Brouard, J.P., Hugonnet, J.E. and Arthur, M. (2003) The CroRS two-component regulatory system is required for intrinsic beta-lactam resistance in *Enterococcus faecalis*, *J Bacteriol*, **185**, 7184-7192.
- Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: A sequence logo generator, *Genome Res*, **14**, 1188-1190.
- Cserzo, M., Eisenhaber, F., Eisenhaber, B. and Simon, I. (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter, *Bioinformatics*, **20**, 136-137.
- Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact, *Trends Biochem Sci*, **23**, 324-328.
- Danhorn, T., Hentzer, M., Givskov, M., Parsek, M.R. and Fuqua, C. (2004) Phosphorus limitation enhances biofilm formation of the plant pathogen *Agrobacterium tumefaciens* through the PhoR-PhoB regulatory system, *J Bacteriol*, **186**, 4492-4501.
- Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) A model of evolutionary change in proteins, *Atlas of Protein Sequence and Structure*, **5**, 345-352.
- Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER, *Nucleic Acids Res*, **27**, 4636-4641.
- Deng, W., Burland, V., Plunkett, G., Boutin, A., Mayhew, G.F., Liss, P., Perna, N.T., Rose, D.J., Mau, B., Zhou, S.G., Schwartz, D.C., Fetherston, J.D., Lindler, L.E.,

- Brubaker, R.R., Plano, G.V., Straley, S.C., McDonough, K.A., Nilles, M.L., Matson, J.S., Blattner, F.R. and Perry, R.D. (2002) Genome sequence of *Yersinia pestis* KIM, *J Bacteriol*, **184**, 4601-4611.
- Deshpande, N., Address, K.J., Bluhm, W.F., Merino-Ott, J.C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z.K., Green, R.K., Flippen-Anderson, J.L., Westbrook, J., Berman, H.M. and Bourne, P.E. (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema, *Nucleic Acids Res*, **33**, D233-D237.
- Djordjevic, S. and Stock, A.M. (1998) Structural analysis of bacterial chemotaxis proteins: Components of a dynamic signaling system, *J Struct Biol*, **124**, 189-200.
- Do, C.B., Mahabhashyam, M.S.P., Brudno, M. and Batzoglou, S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment, *Genome Res*, **15**, 330-340.
- Eddy, S.R. (1998) Profile hidden Markov models, *Bioinformatics*, **14**, 755-763.
- Edgar, R.C. (2004) Local homology recognition and distance measures in linear time using compressed amino acid alphabets, *Nucleic Acids Res*, **32**, 380-385.
- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res*, **32**, 1792-1797.
- Eguchi, Y. and Utsumi, R. (2005) A novel mechanism for connecting bacterial two-component signal-transduction systems, *Trends Biochem Sci*, **30**, 70-72.
- Elsen, S., Swem, L.R., Swem, D.L. and Bauer, C.E. (2004) RegB/RegA, a highly conserved redox-responding global two-component regulatory system, *Microbiol Mol Biol R*, **68**, 263-+.
- Falke, J.J. and Hazelbauer, G.L. (2001) Transmembrane signaling in bacterial chemoreceptors, *Trends Biochem Sci*, **26**, 257-265.
- Feng, D.F., Johnson, M.S. and Doolittle, R.F. (1985) Aligning Amino-Acid Sequences - Comparison of Commonly Used Methods, *J Mol Evol*, **21**, 112-125.

- Fraser, C.M., Eisen, J.A. and Salzberg, S.L. (2000) Microbial genome sequencing, *Nature*, **406**, 799-803.
- Galibert, F., Finan, T.M., Long, S.R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M.J., Becker, A., Boistard, P., Bothe, G., Boutry, M., Bowser, L., Buhrmester, J., Cadieu, E., Capela, D., Chain, P., Cowie, A., Davis, R.W., Dreano, S., Federspiel, N.A., Fisher, R.F., Gloux, S., Godrie, T., Goffeau, A., Golding, B., Gouzy, J., Gurjal, M., Hernandez-Lucas, I., Hong, A., Huizar, L., Hyman, R.W., Jones, T., Kahn, D., Kahn, M.L., Kalman, S., Keating, D.H., Kiss, E., Komp, C., Lalaure, V., Masuy, D., Palm, C., Peck, M.C., Pohl, T.M., Portetelle, D., Purnelle, B., Ramsperger, U., Surzycki, R., Thebault, P., Vandenbol, M., Vorholter, F.J., Weidner, S., Wells, D.H., Wong, K., Yeh, K.C. and Batut, J. (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti*, *Science*, **293**, 668-672.
- Galperin, M.Y. (2005) The Molecular Biology Database Collection: 2005 update, *Nucleic Acids Res*, **33**, D5-D24.
- Galperin, M.Y., Nikolskaya, A.N. and Koonin, E.V. (2001) Novel domains of the prokaryotic two-component signal transduction systems, *Fems Microbiol Lett*, **203**, 11-21.
- Genick, U.K., Soltis, S.M., Kuhn, P., Canestrelli, I.L. and Getzoff, E.D. (1998) Structure at 0.85 angstrom resolution of an early protein photocycle intermediate, *Nature*, **392**, 206-209.
- Georgellis, D., Kwon, O., Lin, E.C.C., Wong, S.M. and Akerley, B.J. (2001) Redox signal transduction by the ArcB sensor kinase of *Haemophilus influenzae* lacking the PAS domain, *J Bacteriol*, **183**, 7206-7212.
- Gong, W.M., Hao, B., Mansy, S.S., Gonzalez, G., Gilles-Gonzalez, M.A. and Chan, M.K. (1998) Structure of a biological oxygen sensor: A new mechanism for heme-driven signal transduction, *P Natl Acad Sci USA*, **95**, 15177-15182.
- Grebe, T.W. and Stock, J.B. (1999) The histidine protein kinase superfamily, *Adv Microb Physiol*, **41**, 139-227.
- Haldimann, A., Fisher, S.L., Daniels, L.L., Walsh, C.T. and Wanner, B.L. (1997) Transcriptional regulation of the *Enterococcus faecium* BM4147 vancomycin resistance gene cluster by the VanS-VanR two-component regulatory system in *Escherichia coli* K-12, *J Bacteriol*, **179**, 5903-5913.

- Harper, S.M., Neil, L.C. and Gardner, K.H. (2003) Structural basis of a phototropin light switch, *Science*, **301**, 1541-1544.
- Hefti, M.H., Francoijs, K.J., de Vries, S.C., Dixon, R. and Vervoort, J. (2004) The PAS fold - A redefinition of the PAS domain based upon structural prediction, *European Journal of Biochemistry*, **271**, 1198-1208.
- Hendrixson, D.R., Akerley, B.J. and DiRita, V.J. (2001) Transposon mutagenesis of *Campylobacter jejuni* identifies a bipartite energy taxis system required for motility, *Mol Microbiol*, **40**, 214-224.
- Henikoff, S. and Henikoff, J.G. (1991) Automated Assembly of Protein Blocks for Database Searching, *Nucleic Acids Res*, **19**, 6565-6572.
- Henikoff, S. and Henikoff, J.G. (1992) Amino-Acid Substitution Matrices from Protein Blocks, *P Natl Acad Sci USA*, **89**, 10915-10919.
- Hoch, J.A. (2000) Two-component and phosphorelay signal transduction, *Curr Opin Microbiol*, **3**, 165-170.
- Hoch, J.A. and Silhavy, T.A. (1995) *Two-component signal transduction*. ASM Press, Washington, D.C.
- Inouye, M. and Dutta, R. (2003) *Histidine Kinases in Signal Transduction*. Academic Press, San Diego, CA.
- Island, M.D. and Kadner, R.J. (1993) Interplay between the Membrane-Associated UhpB and UhpC Regulatory Proteins, *J Bacteriol*, **175**, 5028-5034.
- Jenal, U. (2004) Cyclic di-guanosine-monophosphate comes of age: a novel secondary messenger involved in modulating cell surface structures in bacteria?, *Curr Opin Microbiol*, **7**, 185-191.
- Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol*, **292**, 195-202.
- Kall, L., Krogh, A. and Sonnhammer, E.L.L. (2004) A combined transmembrane topology and signal peptide prediction method, *J Mol Biol*, **338**, 1027-1036.

- Karlin, S. and Ghandour, G. (1985) Multiple-Alphabet Amino-Acid-Sequence Comparisons of the Immunoglobulin K-Chain Constant Domain, *P Natl Acad Sci USA*, **82**, 8597-8601.
- Karp, P.D. (2001) Pathway databases: A case study in computational symbolic theories, *Science*, **293**, 2040-2044.
- Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res*, **30**, 3059-3066.
- Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gill, M. and Karp, P.D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli, *Nucleic Acids Res*, **33**, D334-D337.
- Kimura, M. (1985) *The Neutral Theory of Molecular Evolution*. Cambridge University Press.
- Koh, I.Y.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A. and Rost, B. (2003) EVA: evaluation of protein structure prediction servers, *Nucleic Acids Res*, **31**, 3311-3315.
- Kolb, A., Busby, S., Buc, H., Garges, S. and Adhya, S. (1993) Transcriptional Regulation by Camp and Its Receptor Protein, *Annu Rev Biochem*, **62**, 749-795.
- Konstantinidis, K.T. and Tiedje, J.M. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes, *P Natl Acad Sci USA*, **101**, 3160-3165.
- Koonin, E.V. and Galperin, M.Y. (2003) *Sequence - Evolution - Function : Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers, Boston, MA.
- Koretke, K.K., Lupas, A.N., Warren, P.V., Rosenberg, M. and Brown, J.R. (2000) Evolution of two-component signal transduction, *Mol Biol Evol*, **17**, 1956-1970.

- Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes, *J Mol Biol*, **305**, 567-580.
- Krukonis, E.S. and DiRita, V.J. (2003) From motility to virulence: sensing and responding to environmental signals in *Vibrio cholerae*, *Curr Opin Microbiol*, **6**, 186-190.
- Kumar, S., Tamura, K. and Nei, M. (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment, *Brief Bioinform*, **5**, 150-163.
- Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. and Bork, P. (2004) SMART 4.0: towards genomic data integration, *Nucleic Acids Res*, **32**, D142-D144.
- Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G. and Lu, P.Z. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer, *Science*, **271**, 1247-1254.
- Lipman, D.J., Altschul, S.F. and Kececioglu, J.D. (1989) A Tool for Multiple Sequence Alignment, *P Natl Acad Sci USA*, **86**, 4412-4415.
- Lupas, A. (1997) Predicting coiled-coil regions in proteins, *Curr Opin Struc Biol*, **7**, 388-393.
- Macheroux, P., Hill, S., Austin, S., Eydmann, T., Jones, T., Kim, S.O., Poole, R. and Dixon, R. (1998) Electron donation to the flavoprotein NifL, a redox-sensing transcriptional regulator, *Biochemical Journal*, **332**, 413-419.
- Matsushita, M. and Janda, K.D. (2002) Histidine kinases as targets for new antimicrobial agents, *Bioorgan Med Chem*, **10**, 855-867.
- McCluskey, J., Hinds, J., Husain, S., Witney, A. and Mitchell, T.J. (2004) A two-component system that controls the expression of pneumococcal surface antigen A (PsaA) and regulates virulence and resistance to oxidative stress in *Streptococcus pneumoniae*, *Mol Microbiol*, **51**, 1661-1675.

- Milburn, M.V., Prive, G.G., Milligan, D.L., Scott, W.G., Yeh, J., Jancarik, J., Koshland, D.E. and Kim, S.H. (1991) Three-Dimensional Structures of the Ligand-Binding Domain of the Bacterial Aspartate Receptor with and without a Ligand, *Science*, **254**, 1342-1347.
- Mishra, M., Parise, G., Jackson, K.D., Wozniak, D.J. and Deora, R. (2005) The BvgAS signal transduction system regulates biofilm development in *Bordetella*, *J Bacteriol*, **187**, 1474-1484.
- Mougel, C. and Zhulin, I.B. (2001) CHASE: an extracellular sensing domain common to transmembrane receptors from prokaryotes, lower eukaryotes and plants, *Trends Biochem Sci*, **26**, 582-584.
- Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K. and Pedersen, J.T. (1997) Critical assessment of methods of protein structure prediction (CASP): Round II, *Proteins*, 2-6.
- Needleman, S.B. and Wunsch, C.D. (1970) A General Method Applicable to Search for Similarities in Amino Acid Sequence of 2 Proteins, *J Mol Biol*, **48**, 443-&.
- Nichols, N.N. and Harwood, C.S. (2000) An aerotaxis transducer gene from *Pseudomonas putida*, *Fems Microbiol Lett*, **182**, 177-183.
- Nikolskaya, A.N. and Galperin, M.Y. (2002) A novel type of conserved DNA-binding domain in the transcriptional regulators of the AlgR/AgrA/LytR family, *Nucleic Acids Res*, **30**, 2453-2459.
- Nioche, P., Berka, V., Vipond, J., Minton, N., Tsai, A.L. and Raman, C.S. (2004) Femtomolar sensitivity of a NO sensor from *Clostridium botulinum*, *Science*, **306**, 1550-1553.
- Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J Mol Biol*, **302**, 205-217.
- Novick, R.P. and Jiang, D.R. (2003) The staphylococcal saeRS system coordinates environmental signals with agr quorum sensing, *Microbiol-Sgm*, **149**, 2709-2717.
- Ohki, R., Giyanto, Tateno, K., Masuyama, W., Moriya, S., Kobayashi, K. and Ogasawara, N. (2003) The BceRS two-component regulatory system induces

- expression of the bacitracin transporter, BceAB, in *Bacillus subtilis*, *Mol Microbiol*, **49**, 1135-1144.
- Ottemann, K.M., Xiao, W.Z., Shin, Y.K. and Koshland, D.E. (1999) A piston model for transmembrane signaling of the aspartate receptor, *Science*, **285**, 1751-1754.
- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) The use of gene clusters to infer functional coupling, *P Natl Acad Sci USA*, **96**, 2896-2901.
- Pappas, K.M., Weingart, C.L. and Winans, S.C. (2004) Chemical communication in proteobacteria: biochemical and structural studies of signal synthases and receptors required for intercellular signalling, *Mol Microbiol*, **53**, 755-769.
- Parkinson, J.S. (1993) Signal-Transduction Schemes of Bacteria, *Cell*, **73**, 857-871.
- Parkinson, J.S. and Kofoed, E.C. (1992) Communication Modules in Bacterial Signaling Proteins, *Annu Rev Genet*, **26**, 71-112.
- Paul, R., Weiser, S., Amiot, N.C., Chan, C., Schirmer, T., Giese, B. and Jenal, U. (2004) Cell cycle-dependent dynamic localization of a bacterial response regulator with a novel di-guanylate cyclase output domain, *Gene Dev*, **18**, 715-727.
- Peach, M.L., Hazelbauer, G.L. and Lybrand, T.P. (2002) Modeling the transmembrane domain of bacterial chemoreceptors, *Protein Sci*, **11**, 912-923.
- Pearson, W.R. and Lipman, D.J. (1988) Improved Tools for Biological Sequence Comparison, *P Natl Acad Sci USA*, **85**, 2444-2448.
- Pei, J.M. and Grishin, N.V. (2001) GGDEF domain is homologous to adenylyl cyclase, *Proteins*, **42**, 210-216.
- Pei, J.M., Sadreyev, R. and Grishin, N.V. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency, *Bioinformatics*, **19**, 427-428.
- Pellequer, J.L., Brudler, R. and Getzoff, E.D. (1999) Biological sensors: More than one way to sense oxygen, *Current Biology*, **9**, R416-R418.

- Pellequer, J.L., Wager-Smith, K.A., Kay, S.A. and Getzoff, E.D. (1998) Photoactive yellow protein: A structural prototype for the three-dimensional fold of the PAS domain superfamily, *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5884-5890.
- Ponting, C.P. and Aravind, L. (1997) PAS: a multifunctional domain family comes to light, *Current Biology*, **7**, R674-R677.
- Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res*, **33**, D501-D504.
- Read, T.D., Peterson, S.N., Tourasse, N., Baillie, L.W., Paulsen, I.T., Nelson, K.E., Tettelin, H., Fouts, D.E., Eisen, J.A., Gill, S.R., Holtzapple, E.K., Okstad, O.A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J.F., Beanan, M.J., Dodson, R.J., Brinkac, L.M., Gwinn, M., DeBoy, R.T., Madpu, R., Daugherty, S.C., Durkin, A.S., Haft, D.H., Nelson, W.C., Peterson, J.D., Pop, M., Khouri, H.M., Radune, D., Benton, J.L., Mahamoud, Y., Jiang, L.X., Hance, I.R., Weidman, J.F., Berry, K.J., Plaut, R.D., Wolf, A.M., Watkins, K.L., Nierman, W.C., Hazen, A., Cline, R., Redmond, C., Thwaite, J.E., White, O., Salzberg, S.L., Thomason, B., Friedlander, A.M., Koehler, T.M., Hanna, P.C., Kolsto, A.B. and Fraser, C.M. (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria, *Nature*, **423**, 81-86.
- Rebbapragada, A., Johnson, M.S., Harding, G.P., Zuccarelli, A.J., Fletcher, H.M., Zhulin, I.B. and Taylor, B.L. (1997) The Aer protein and the serine chemoreceptor Tsr independently sense intracellular energy levels and transduce oxygen, redox, and energy signals for *Escherichia coli* behavior, *P Natl Acad Sci USA*, **94**, 10541-10546.
- Reid, C.J. and Poole, P.S. (1998) Roles of DctA and DctB in signal detection by the dicarboxylic acid transport system of *Rhizobium leguminosarum*, *J Bacteriol*, **180**, 2660-2669.
- Reinelt, S., Hofmann, E., Gerharz, T., Bott, M. and Madden, D.R. (2003) The structure of the periplasmic ligand-binding domain of the sensor kinase CitA reveals the first extracellular PAS domain, *J Biol Chem*, **278**, 39189-39196.
- Repik, A., Rebbapragada, A., Johnson, M.S., Haznedar, J.O., Zhulin, I.B. and Taylor, B.L. (2000) PAS domain residues involved in signal transduction by the Aer redox sensor of *Escherichia coli*, *Molecular Microbiology*, **36**, 806-816.

- Rogers, Y.H. and Venter, J.C. (2005) Genomics - Massively parallel sequencing, *Nature*, **437**, 326-327.
- Rost, B., Sander, C. and Schneider, R. (1994) Redefining the Goals of Protein Secondary Structure Prediction, *J Mol Biol*, **235**, 13-26.
- Saitou, N. and Nei, M. (1987) The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees, *Mol Biol Evol*, **4**, 406-425.
- Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res*, **29**, 2994-3005.
- Schuster, M., Silversmith, R.E. and Bourret, R.B. (2001) Conformational coupling in the chemotaxis response regulator CheY, *P Natl Acad Sci USA*, **98**, 6003-6008.
- Shelver, D., Kerby, R.L., He, Y.P. and Roberts, G.P. (1997) CooA, a CO-sensing transcription factor from *Rhodospirillum rubrum*, is a CO-binding heme protein, *P Natl Acad Sci USA*, **94**, 11216-11220.
- Shu, C.J., Ulrich, L.E. and Zhulin, I.B. (2003) The NIT domain: a predicted nitrate-responsive module in bacterial sensory receptors, *Trends Biochem Sci*, **28**, 121-124.
- Shu, C.Y.J. and Zhulin, I.B. (2002) ANTAR: an RNA-binding domain in transcription antitermination regulatory proteins, *Trends Biochem Sci*, **27**, 3-5.
- Smith, T.F. and Waterman, M.S. (1981) Identification of Common Molecular Subsequences, *J Mol Biol*, **147**, 195-197.
- Sourjik, V. (2004) Receptor clustering and signal processing in E coli chemotaxis, *Trends Microbiol*, **12**, 569-576.
- Spiro, S. and Guest, J.R. (1990) Fnr and Its Role in Oxygen-Regulated Gene-Expression in *Escherichia-Coli*, *Fems Microbiol Rev*, **75**, 399-428.

- Staal, M., Meysman, F.J.R. and Stal, L.J. (2003) Temperature excludes N₂-fixing heterocystous cyanobacteria in the tropical oceans, *Nature*, **425**, 504-507.
- Stephenson, K. and Hoch, J.A. (2002) Evolution of signalling in the sporulation phosphorelay, *Mol Microbiol*, **46**, 297-304.
- Stephenson, K. and Hoch, J.A. (2002) Two-component and phosphorelay signal-transduction systems as therapeutic targets, *Curr Opin Pharmacol*, **2**, 507-512.
- Stock, A.M., Robinson, V.L. and Goudreau, P.N. (2000) Two-component signal transduction, *Annu Rev Biochem*, **69**, 183-215.
- Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A. (2003) The COG database: an updated version includes eukaryotes, *Bmc Bioinformatics*, **4**, -.
- Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families, *Science*, **278**, 631-637.
- Taylor, B.L. and Zhulin, I.B. (1999) PAS domains: Internal sensors of oxygen, redox potential, and light, *Microbiology and Molecular Biology Reviews*, **63**, 479-506.
- Taylor, B.L. and Zhulin, I.B. (1999) PAS domains: Internal sensors of oxygen, redox potential, and light, *Microbiol Mol Biol R*, **63**, 479-+.
- Taylor, B.L., Zhulin, I.B. and Johnson, M.S. (1999) Aerotaxis and other energy-sensing behavior in bacteria, *Annu Rev Microbiol*, **53**, 103-128.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice, *Nucleic Acids Res*, **22**, 4673-4680.
- Ulijasz, A.T., Andes, D.R., Glasner, J.D. and Weisblum, B. (2004) Regulation of iron transport in *Streptococcus pneumoniae* by RitR, an orphan response regulator, *J Bacteriol*, **186**, 8123-8136.

- Ulrich, L.E., Koonin, E.V. and Zhulin, I.B. (2005) One-component systems dominate signal transduction in prokaryotes, *Trends Microbiol*, **13**, 52-56.
- Ulrich, L.E. and Zhulin, I.B. (2005) Four-Helix Bundle: a Ubiquitous Sensory Module in Prokaryotic Signal Transduction, *Bioinformatics*, in press.
- Van Nimwegen, E. (2003) Scaling laws in the functional content of genomes, *Trends Genet*, **19**, 479-484.
- Vannini, A., Volpari, C., Gargioli, C., Muraglia, E., Cortese, R., De Francesco, R., Neddermann, P. and Di Marco, S. (2002) The crystal structure of the quorum sensing protein TraR bound to its autoinducer and target DNA, *Embo J*, **21**, 4393-4401.
- von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P. and Snel, B. (2003) STRING: a database of predicted functional associations between proteins, *Nucleic Acids Res*, **31**, 258-261.
- Walderhaug, M.O., Polarek, J.W., Voelkner, P., Daniel, J.M., Hesse, J.E., Altendorf, K. and Epstein, W. (1992) Kdpd and Kdpe, Proteins That Control Expression of the Kdpabc Operon, Are Members of the 2-Component Sensor-Effector Class of Regulators, *J Bacteriol*, **174**, 2152-2159.
- Weiss, V., Kramer, G., Dunnebier, T. and Flotho, A. (2002) Mechanism of regulation of the bifunctional histidine kinase NtrB in Escherichia coli, *J Mol Microb Biotech*, **4**, 229-233.
- Wootton, J.C. and Federhen, S. (1993) Statistics of Local Complexity in Amino-Acid-Sequences and Sequence Databases, *Comput Chem*, **17**, 149-163.
- Zahrt, T.C. and Deretic, V. (2001) Mycobacterium tuberculosis signal transduction system required for persistent infections, *P Natl Acad Sci USA*, **98**, 12706-12711.
- Zhang, H.Z., Hackbarth, C.J., Chansky, K.M. and Chambers, H.F. (2001) A proteolytic transmembrane signaling pathway and resistance to beta-lactams in staphylococci, *Science*, **291**, 1962-1965.

- Zhao, R., Collins, E.J., Bourret, R.B. and Silversmith, R.E. (2002) Structure and catalytic mechanism of the E-coli chemotaxis phosphatase CheZ, *Nat Struct Biol*, **9**, 570-575.
- Zhou, H. and Zhou, Y. (2005) SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures, *Bioinformatics*, **21**, 3615-3621.
- Zhulin, I.B. (2001) The superfamily of chemotaxis transducers: From physiology to genomics and back, *Advances in Microbial Physiology*, Vol 45, **45**, 157-198.
- Zhulin, I.B., Nikolskaya, A.N. and Galperin, M.Y. (2003) Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in Bacteria and Archaea, *J Bacteriol*, **185**, 285-294.
- Zhulin, I.B. and Taylor, B.L. (1997) PAS domain S-boxes in Archaea, bacteria and sensors for oxygen and redox, *Trends in Biochemical Sciences*, **22**, 331-333.
- Zimmer, D.P., Soupene, E., Lee, H.L., Wendisch, V.F., Khodursky, A.B., Peter, B.J., Bender, R.A. and Kustu, S. (2000) Nitrogen regulatory protein C-controlled genes of Escherichia coli: Scavenging as a defense against nitrogen limitation, *P Natl Acad Sci USA*, **97**, 14674-14679.